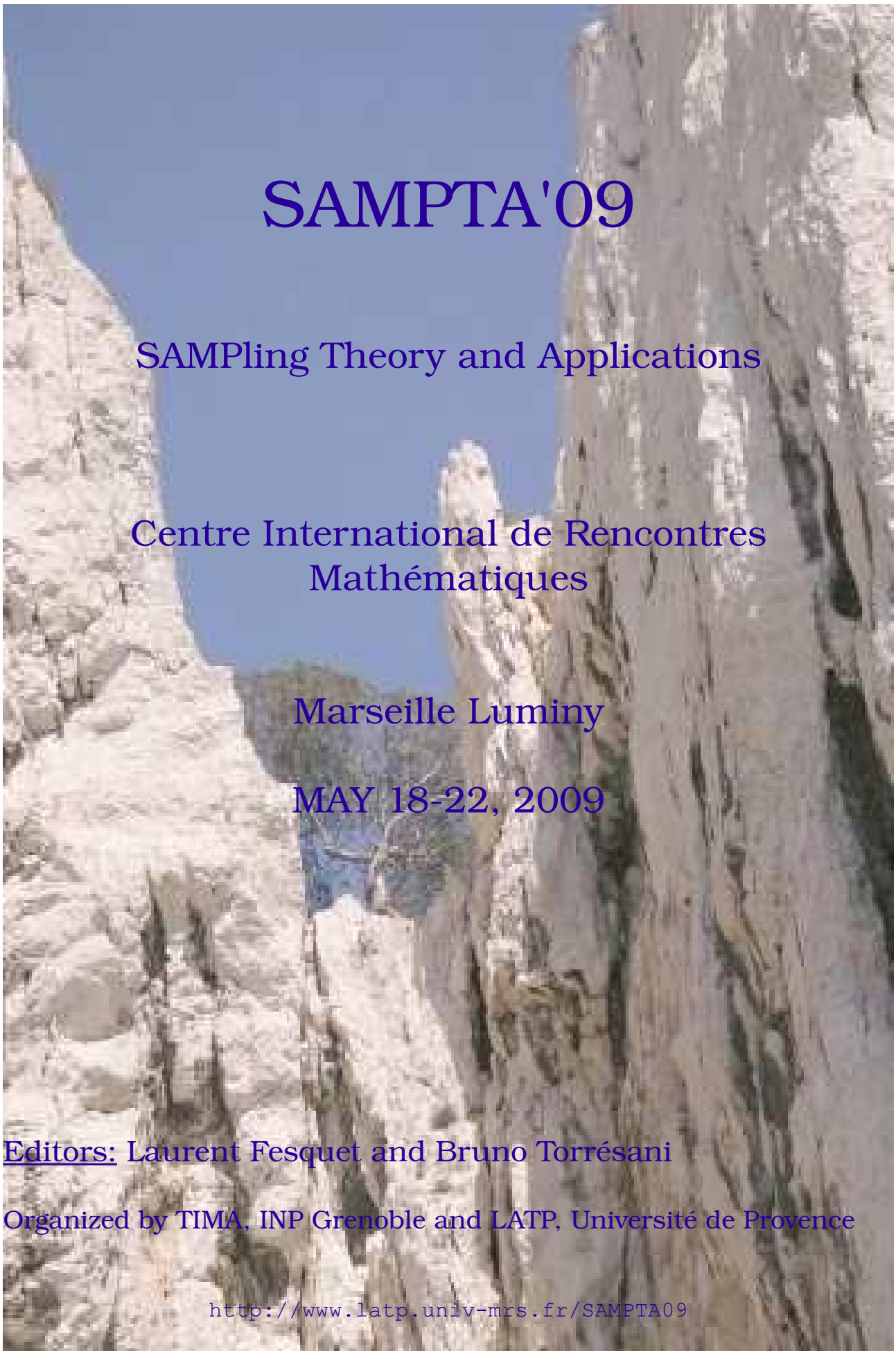
SAMPTA'09, International Conference on SAMPling Theory and Applications

Laurent Fesquet, Bruno Torrésani



SAMPTA'09

SAMPling Theory and Applications

Centre International de Rencontres Mathématiques

Marseille Luminy

MAY 18-22, 2009

Editors: Laurent Fesquet and Bruno Torrésani

Organized by TIMA, INP Grenoble and LATP, Université de Provence

http://www.latp.univ-mrs.fr/SAMPTA09

SAMPTA'09 Participants

SAMPTA'09, the 8th international conference on Sampling Theory and Applications, was organized in Marseille-Luminy, on May 18-22, 2009. The previous conferences were held in Riga (Latvia) in 1995, Aveiro (Portugal) in 1997, Loen (Norway) in 1999, Orlando (USA) in 2001, Salzburg (Austria) in 2003, Samsun (Turkey) in 2005 and Thessaloniki (Greece) in 2007.

The purpose of SAMPTA's is to bring together mathematicians and engineers interested in sampling theory and its applications to related fields (such as signal and image processing, coding theory, control theory, complex analysis, harmonic analysis, differential equations) to exchange recent advances and to discuss open problems.

SAMPTA09 gathered around 160 participants from various countries and scientific areas, The conference benefited from the infrastructure of CIRM, the *Centre International de Rencontres Mathématiques*, an institute mainly sponsored by the french *Centre National de la Recherche Scientifique* (CNRS) and the *French Mathematical Society* (SMF).

Organizing committee:

General chairs

- Laurent Fesquet (TIMA, Grenoble Institute of Technology)
- Bruno Torrésani (LATP, Université de Provence, Marseille)

Local committee

- Sandrine Anthoine (I3S, CNRS, Sophia-Antipolis)
- Karim Kellay (LATP, Université de Provence, Marseille)
- Matthieu Kowalski (LATP, Université de Provence, Marseille)
- Clothilde Melot (LATP, Université de Provence, Marseille)
- El Hassan Youssfi (LATP, Université de Provence, Marseille

Program committee:

Program chairs

- Yonina Eldar (Electrical Engineering, Technion, Israel Institute of Technology)
- Karlheinz Gröchenig (Fakultät für Mathematik, University of Vienna)
- Sinan Gunturk (Mathematics Department, Courant Institute, New York)
- Michael Unser (Biomedical Imaging Group, EPFL Lausanne)

Special sessions organizers

- Bernhard Bodmann (Dept of Mathematics, University of Houston, USA)
- Pierluigi Dragotti (Dept of Electronic and Alectrical Engineering, Imperial College London, UK)
- Yonina Eldar (Electrical Engineering, Technion, Israel Institute of Technology, Israel)
- Laurent Fesquet (TIMA, INP Grenoble, France)
- Massimo Fornasier (RICAM, Linz University, Austria)
- Hakan Johansson (Dept of Electrical Engineering, Linkoping University, Sweden)
- Gitta Kutyniok (Universität Osnabrück, Germany)
- Pina Marziliano (Nanyang Technological University, Singapore)
- Götz Pfander (International University Bremen, Germany)
- Hölger Rauhut (Hausdorff Center for Mathematics, University of Bonn, Germany)
- Jared Tanner (School of Mathematics, University of Edimburgh, Scotland)
- Christian Vogel (ISI, ETH Zurich, Switzerland)
- Ozgur Yilmaz (Dept of Mathematics, University of British Columbia, Vancouver, Canada)

Acknowledgements

The conference was extremely successful, thanks mainly to the participants, whose scientific contributions were remarkably good. Thanks are also due to the members of the organizing committee and the program committee, as well as all the reviewers who participated in the selection of contributions.

We would also like to thank the CIRM staff for the practical organization (accomodation, conference facilities,...) and their constant availability, and the Faculté des Sciences de Luminy for lending a conference room for the plenary sessions.

The secretary staff at LATP was instrumental in all aspects of the organization, from the scientific part to the social events.

Finally, we would like to thank the sponsors of the conference: *CIRM* (CNRS and French Mathematical Society), Université de Provence, the European Excellence Center for Time-Frequency Analysis (*EUCETIFA*), the City of Marseille and the Conseil Général des Bouches du Rhône for their financial support.

SampTA Technical Program
Monday 18 May 2009

09:10 - 09:30 **Opening Session – Amphi 8**
09:30 - 10:30 **Plenary talk - Amphi 8 – Chair: K. Gröchenig**
Gabor frames in Complex Analysis , Yura Lyubarskii
10:30 - 11:00 **Coffee break**
11:00 - 12:00 **Plenary talk - Amphi 8 – Chair: K. Gröchenig**
A Prior-Free Approach to Signal and Image Denoising: the SURE-LET Methodology, Thierry Blu
12:00 - 14:00 **Lunch**
14:00 - 16:00 **Special session – Auditorium**
Sparse approximation and high-dimensional geometry - Chair: J. Tanner
#192. Dense Error Correction via L1-Minimization, John Wright, Yi Ma
#204. Recovery of Clustered Sparse Signals from Compressive Measurements, Volkan Cevher, Piotr Indyk, Chinmay Hegde, Richard G. Baraniuk
#206. Sparse Recovery via l_q-minimization for $0 < q \leq 1$, Simon Foucart
#210. The Balancedness Properties of Linear Subpaces and Signal Recovery Robustness in Compressive Sensing, Weiyu Xu
#207. Phase Transitions Phenomena in Compressed Sensing, Jared Tanner
#167 Optimal Non-Linear Models, Akram Aldroubi, Carlos Cabrelli,Ursula Molter
14:00 - 16:00 **General session – room 1**
General sampling - Chair: Y. Lyubarskii
#78. Linear Signal Reconstruction from Jittered Sampling, Alessandro Nordio, Carla-Fabiana Chiasserini, Emanuele Viterbo
#81. Zero-two derivative sampling, Gerhard Schmeisser
#115. On average sampling restoration of Piranashvili-type harmonizable processes, Andriy Ya. Olenko, Tibor K. Pogany
#86. Uniform Sampling and Reconstruction of Trivariate Functions, Alireza Entezari
#184. On Subordination Principles for Generalized Shannon Sampling Series, Andi Kivinukk and Gert Tamberg
#104. The Class of Bandlimited Functions with Unstable Reconstruction under Thresholding, Holger Boche, Ullrich J. Mönich
16:00 - 16:30 **Coffee break**
16:30 – 18:30 **Special Session – Auditorium**
Compressed sensing - Chair: Y. Eldar
#150. Sampling Shift-Invariant Signals with Finite Rate of Innovation, Kfir Gedalyahu, Yonina C. Eldar
#105. Compressed sensing signal models - to infinity and beyond?, Thomas Blumensath, Mike Davies
#110. Compressed sampling Via Huffman Codes, Akram Aldroubi, Haichao Wang, Kourosh Zaringhalam
#126. On Lp minimisation, instance optimality, and restricted isometry constants for sparse approximation, Michael Davies, Rémi Gribonval
#73. Signal recovery from incomplete and inaccurate measurements via ROMP, Deanna Needell, Roman Vershynin
#169 Sparse approximation and the MAP, Akram Aldroubi, Romain Tessera
16:30 – 18:30 **Special Session – room 1**
Frame theory and oversampling - Chair: B. Bodmann
#118. Invariance of Shift Invariance Spaces, Akram Aldroubi, Carlos Cabrelli, Christopher Heil, Keri Kornelson, Ursula Molter
#188. Gabor frames with reduced redundancy, Ole Christensen, Hong Oh Kim, Rae Young Kim
#141. Gradient descent of the frame potential, Peter G. Casazza, Matthew Fickus
#201. Error Correction for Erasures of Quantized Frame Coefficients, Bernhard G. Bodmann, Peter G. Casazza ,Gitta Kutyniok, Steven Senger
#199. Linear independence and coherence of Gabor systems in finite dimensional spaces, Götz E. Pfander

Tuesday 19 May 2009

09:10 - 10:30 **Special Session – Auditorium**

Efficient design and implementation of sampling rate conversion, resampling and signal reconstruction methods - Chair: H. Johansson and C. Vogel

#194. Efficient design and implementation of sampling rate conversion, resampling, and signal reconstruction methods, Håkan Johansson, Christian Vogel

#171. Structures for Interpolation, Decimation, and Nonuniform Sampling Based on Newton's Interpolation Formula, Vesa Lehtinen, Markku Renfors

#79. Chromatic Derivatives, Chromatic Expansions and Associated Function Spaces, Aleksandar Ignjatovic

#84. Estimation of the Length and the Polynomial Order of Polynomial-based Filters, Djordje Babic, Heinz G. Göckler

09:10 - 10:30 **General Session – room 1**

Time frequency and frames - Chair: J.P. Antoine

#121. An Efficient Algorithm for the Discrete Gabor Transform using full length Windows, Peter L. Søndergaard

#82. Matrix Representation of Bounded Linear Operators By Bessel Sequences, Frames and Riesz Sequence, Peter Balazs

#124. Nonstationary Gabor Frames, Florent Jaillet, Peter Balazs, Monika Dörfler

#140. A Nonlinear Reconstruction Algorithm from Absolute Value of Frame Coefficients for Low Redundancy Frames, Radu Balan

10:30 - 11:00 **Coffee break**

11:00 - 12:00 **Plenary talk – Amphi 8 – Chair: A. Aldroubi**

Harmonic and multiscale analysis of and on data sets in high-dimensions, Mauro Maggioni

12:00 - 14:00 **Lunch**

14:00 - 16:00 **Special session – Auditorium**

Geometric multiscale analysis I - Chair: G. Kutyniok

#74. Analysis of Singularities and Edge Detection using the Shearlet Transform, Glenn Easley, Kanghui Guo, Demetrio Labate

#98. Discrete Shearlet Transform: New Multiscale Directional Image Representation, Wang-Q Lim

#125. The Continuous Shearlet Transform in Arbitrary Space Dimensions, Frame Construction, and Analysis of singularities, Stephan Dahlke, Gabriele Steidl, Gerd Teschke

#193. Computable Fourier Conditions for Alias-Free Sampling and Critical Sampling, Yue M. Lu, Minh N. Do, Richard S. Laugesen

#149. Compressive-wavefield simulations, Felix J. Herrmann, Yogi Erlangga, Tim T. Y. Lin

#164. Analysis of Singularity Lines by Transforms with Parabolic Scaling, Panuvuth Lakhonchai, Jouni Sampo, Songkiat Sumetkijakan

14:00 - 16:00 **Special session – room 1**

Sampling and communication - Chair: G. Pfander

#146. Erasure-proof coding with fusion frames, Bernhard G. Bodmann, Gitta Kutyniok, Ali Pezeshki

#175. Operator Identification and Sampling, Götz Pfander, David Walnut

#116. A Kashin Approach to the Capacity of the Discrete Amplitude Constrained Gaussian Channel, Brendan Farrell, Peter Jung

#147. Irregular and Multi-channel sampling in Operator Paley-Wiener spaces, Yoon Mi Hong, G. Pfander

#136. Low-rate Wideband Receiver, Moshe Mishali, Yonina Eldar

#151. Representation of operators by sampling in the time-frequency domain, Monika Dörfler, Bruno Torrésani

16:00 - 16:30 **Coffee break**

16:30 - 17:30 **Special session – Auditorium**

Geometric multiscale analysis II - Chair: G. Kutyniok

#120. Geometric Wavelets for Image Processing: Metric Curvature of Wavelets, Emil Saucan, Chen Sagiv, Eli Appleboim

#102. Image Approximation by Adaptive Tetrolet Transform, Jens Krommweh

#202. Geometric Separation using a Wavelet-Shearlet Dictionary, David L. Donoho, Gitta Kutyniok

16:30 - 17:30 **General session – room 1**
Sparsity and compressed sensing - Chair: R. Gribonval
#127. Sparse Coding in Mass Spectrometry, Stefan Schiffler, Dirk Lorenz,Theodore Alexandrov
#161.Quasi-Random Sequences for Signal Sampling and Recovery, Miroslaw Pawlak, Ewaryst Rafajlowicz
17:30 - 18:30 **Poster session**
#75. Sparse representation with harmonic wavelets, Carlo Cattani
#85. Reconstruction of signals in a shift-invariant space from nonuniform samples, Junxi Zhao
#92. Spline Interpolation in Piecewise Constant Tension, Masaru Kamada, Rentsen Enkhbat#95. The Effect of Sampling Frequency on a FFT Based Spectral Estimator, Saeed Ayat
#99. Nonlinear Locally Adaptive Wavelet Filter Banks, Gerlind Plonka and Stefanie Tenorth
#111. Continuous Fast Fourier Sampling, Praveen K. Yenduri, Anna C. Gilbert
#134. Double Dirichlet averages and complex B-splines, Peter Massopust
#135. Sampling in cylindrical 2D PET, Yannick Grondin, Laurent Desbat, Michel Desvignes
#148. Significant Reduction of Gibbs' Overshoot with Generalized Sampling Method, Yufang Hao, Achim Kempf
#156. Optimized Sampling Patterns for Practical Compressed MRI, Muhammad Usman, Philip G. Batchelor
#160. A study on sparse signal reconstruction from interlaced samples by l1-norm minimization, Akira Hirabayashi
#162. Multiresolution analysis on multidimensional dyadic grids, Douglas A. Castro, Sônia M. Gomes, Anamaria Gomide, Andrielber S. Oliveira, Jorge Stolfi
#165. Adaptive and Ultra-Wideband Sampling via Signal Segmentation and Projection, Stephen D. Casey, Brian M. Sadler
#174. Non-Uniform Sampling Methods for MRI, Steven Troxler
#187. On approximation properties of sampling operators defined by dilated kernels, Andi Kivinukk, Gert Tamberg

Wednesday 20 May 2009

09:10 - 10:30 **Special Session – Auditorium**
Sampling and industrial applications - Chair: L. Fesquet
#182. A coherent sampling-based method for estimating the jitter used as entropy source for True Random Number Generators, Boyan Valtchanov, Viktor Fischer, Alain Aubert
#91. Orthogonal exponential spline pulses with application to impulse radio, Masaru Kamada, Semih Özlem, Hiromasa Habuchi
#117. Effective Resolution of an Adaptive Rate ADC, Saeed Mian Qaisar, Laurent Fesquet, Marc Renaudin
#157. An Event-Based PID Controller With Low Computational Cost, Sylvain Durand, Nicolas Marchand
09:10 -10:30 **General Session – room 1**
Wavelets, multiresolution and multirate sampling – Chair: D. Walnut
#189. Asymmetric Multi-channel Sampling in Shift Invariant Spaces, Sinuk Kang,Kil Hyun Kwon
#89. Sparse Data Representation on the Sphere using the Easy Path Wavelet Transform, Gerlind Plonka, Daniela Rosca
#114. On the incoherence of noiselet and Haar bases, Tomas Tuma, Paul Hurley
#138. Adaptive compressed image sensing based on wavelet modeling and direct sampling, Shay Deutsch, Amir Averbuch, Shai Dekel10:30 - 11:00 **Coffee break**
11:00 - 12:00 **Plenary talk – Amphi 6 – Chair: G. Teschke**
Recent Developments in Iterative Shrinkage/Thresholding Algorithms, Mario Figueiredo
12:00 - 14:00 **Lunch**
14:00 - 23:00 **Social event**

Thursday 21 May 2009

09:10 - 10:30 **General Session – Auditorium**
Adaptive techniques – Chair: N. Marchand
#68. Adaptive transmission for lossless image reconstruction, Elisabeth Lahalle, Gilles Fleury, Rawad

Zgheib
#129. A fully non-uniform approach to FIR filtering, Brigitte Bidegaray-Fesquet, Laurent Fesquet
#72. Sampling of bandlimited functions on combinatorial graphs, Isaac Pesenson, Meyer Pesenson
#01. Pseudospectral Fourier reconstruction with the inverse polynomial reconstruction method, Karlheinz Groechenig, Tomasz Hrycak
09:10 - 10:30 **General Session – room 1**
General sampling – Chair: A. Jerri
#112. Geometric Sampling of Images, Vector Quantization and Zador's Theorem, Emil Saucan, Eli Appleboim, Yehoshua Y. Zeevi
#168. On sampling lattices with similarity scaling relationships, Steven Bergner, Dimitri Van De Ville, Thierry Blu, Torsten Möller
#83. Scattering Theory and Sampling of Bandlimited Hardy Space Functions, Ahmed I. Zayed, Marianna Shubov
#119. Sampling of Homogeneous Polynomials, Somantika Datta, Stephen D. Howard, Douglas Cochran
10:30 - 11:00 **Coffee break**
11:00 - 12:00 **Plenary talk – Amphi 6 – Chair: S. Güntürk**
A Taste of Compressed Sensing, Ron DeVore
12:00 - 14:00 **Lunch**
14:00 - 16:00 **Special Session – Auditorium**
Sampling using finite rate of innovation principles I - Chair: P. Dragotti and P. Marziliano
#113. The Generalized Annihilation Property --- A Tool For Solving Finite Rate of Innovation Problems, Thierry Blu
#100. Sampling of Sparse Signals in Fractional Fourier Domain, Ayush Bhandari, Pina Marziliano
#153. A method for generalized sampling and reconstruction of finite-rate-of-innovation signals, Chandra Sekhar Seelamantula, Michael Unser
#80. An ``algebraic'' reconstruction of piecewise-smooth functions from integral measurements, Dima Batenkov, Niv Sarig, Yosef Yomdin
#108. Estimating Signals With Finite Rate of Innovation From Noisy Samples: A Stochastic Algorithm, Vincent Y. F. Tan, Vivek K. Goyal
14:00 - 16:00 **Special Session – room 1**
Mathematical aspects of compressed sensing - Chair: H. Rauhut
#195. Orthogonal Matching Pursuit with random dictionaries, Paweł Bechler, Przemysław Wojtaszczyk
#178. A short note on nonconvex compressed sensing, Rayan Saab, Ozgur Yilmaz
#190. Domain decomposition methods for compressed sensing, Massimo Fornasier, Andreas Langer, Carola-Bibiane Schönlieb
#197. Free discontinuity problems meet iterative thresholding, Rachel Ward, Massimo Fornasier
#198. Concise Models for Multi-Signal Compressive Sensing, Mike Wakin
#196. Average case analysis of multichannel Basis Pursuit, Yonina Eldar, Holger Rauhut
16:00 - 16:30 **Coffee break**
16:30 - 17:30 **Special session – Auditorium**
Sampling using finite rate of innovation principles II - Chair: P. Dragotti and P. Marziliano
#96. Distributed Sensing of Signals Under a Sparse Filtering Model, Ali Hormati, Olivier Roy, Yue M. Lu, Martin Vetterli
#154. Multichannel Sampling of Translated, Rotated and Scaled Bilevel Polygons Using Exponential Splines, Hojjat Akhondi Asl, Pier Luigi Dragotti
16:30 - 17:30 **General session – room 1**
Signal Analysis and compressed sensing – Chair: A. Cohen
#176. General Perturbations of Sparse Signals in Compressed Sensing, Matthew Herman, Thomas Strohmer
#203. Limits of Deterministic Compressed Sensing Considering Arbitrary Orthonormal Basis for Sparsity, Arash Amini, Farokh Marvasti
#185. Analysis of High-Dimensional Signal Data by Manifold Learning and Convolutions, Mijail Guillemard, Armin Iske
17:30 - 18:30 Poster session

See Tuesday poster session.

Friday 22 May 2009

09:10 - 10:30 **General Session – Auditorium**

Kernels and unusual Paley-Wiener spaces – Chair: G. Schmeisser

#131. Geometric Reproducing Kernels for Signal Reconstruction, Eli Appleboim, Emil Saucan, Yehoshua Y. Zeevi

#137. Concrete and discrete operator reproducing formulae for abstract Paley-Wiener space, John R. Higgins

#132. Multivariate Complex B-Splines, Dirichlet Averages and Difference Operators, Brigitte Forster, Peter Massopust

#143. Explicit localization estimates for spline-type spaces, José Luis Romero

09:10 - 10:30 **General Session – room 1**

Reconstruction, time and frequency analysis - Chair: R. Balan

#70. Daubechies Localization Operator in Bargmann-Fock Space and Generating Function of Eigenvalues of Localization Operator, Kunio Yoshino

#97 Optimal Characteristic of Optical Filter for White Light Interferometry based on Sampling Theory, Hidemitsu Ogawa and Akira Hirabayashi

#90. Signal-dependent sampling and reconstruction method of signals with time-varying bandwidth, Modris Greitans , Rolands Shavelis

#177. A Fast Fourier Transform with Rectangular Output on the BCC and FCC Lattices, Usman Raza Alim, Torsten Moeller

10:30 - 11:00 **Coffee break**

11:00 - 12:00 **Plenary talk – Amphi 6**

Compressed Sensing in Astronomy, Jean-Luc Starck

12:00 - 14:00 **Lunch**

14:00 - 16:00 **Special Session – Auditorium**

Sampling and quantization - Chair: O. Yilmaz

#180. Finite Range Scalar Quantization for Compressive Sensing, Jason N. Laska, Petros Boufounos, Richard G. Baraniuk

#107. Quantization for Compressed Sensing Reconstruction, John Z. Sun, Vivek K Goyal

#106. Determination of Idle Tones in Sigma-Delta Modulation by Ergodic Theory, Nguyen T. Thao

#172. Noncanonical reconstruction for quantized frame coefficients, Alexander M. Powell

#166. Stability Analysis of Sigma-Delta Quantization Schemes with Linear Quantizers, Percy Deift, Sinan Güntürk, Felix Krahmer

14:00 - 16:00 **Special Session – room 1**

Sampling and inpainting - Chair: M. Fornasier

#191. Image Inpainting Using a Fourth-Order Total Variation Flow, Carola-Bibiane Schönlieb, Andrea Bertozzi, Martin Burger, Lin He

#139. Reproducing kernels and colorization, Minh Q. Ha, Sung Ha Kang, Triet M. Le

#123. Edge Orientation Using Contour Stencils, Pascal Getreuer

#71. Image Segmentation Through Efficient Boundary Sampling, Alex Chen, Todd Wittman, Alexander Tartakovsky, Andrea Bertozzi

#103. Report on Digital Image Processing for Art Historians, Bruno Cornelis, Ann Dooms, Ingrid Daubechies, Peter Schelkens

#158. Smoothing techniques for convex problems. Applications in image processing, Pierre Weiss, Mikael Carlavan, Laure Blanc-Féraud, Josiane Zerubia

SAMPTA'09

Special Sessions

Special session on

Sparse approximation and high-dimensional geometry

Chair: Jared TANNER

Recovery of Clustered Sparse Signals from Compressive Measurements

Volkan Cevher(1), Piotr Indyk(1,2), Chinmay Hegde(1), and Richard G. Baraniuk(1)
(1) Electrical and Computer Engineering, Rice University, Houston, TX
(2) Computer Science and Artificial Intelligence Lab, MIT, Cambridge, MA

Abstract:

We introduce a new signal model, called (K, C)-sparse, to capture K-sparse signals in N dimensions whose nonzero coefficients are contained within at most C clusters, with $C < K \ll N$. In contrast to the existing work in the sparse approximation and compressive sensing literature on block sparsity, no prior knowledge of the locations and sizes of the clusters is assumed. We prove that $\mathcal{O}\left(K + C\log(N/C)\right))$ random projections are sufficient for (K, C)-model sparse signal recovery based on subspace enumeration. We also provide a robust polynomial-time recovery algorithm for (K, C)-model sparse signals with provable estimation guarantees.

1. Introduction

Compressive sensing (CS) is an alternative to Shannon/Nyquist sampling for the acquisition of sparse or compressible signals in an appropriate basis [1, 2]. By sparse, we mean that only K of the N basis coefficients are nonzero, where $K \ll N$. By compressible, we mean the basis coefficients, when sorted, decay rapidly enough to zero so that they can be well-approximated as K-sparse. Instead of taking periodic samples of a signal, CS measures inner products with random vectors and then recovers the signal via a sparsity-seeking convex optimization or greedy algorithm. The number of compressive measurements M necessary to recover a sparse signal under this framework grows as $M = \mathcal{O}\left(K \log(N/K)\right)$

In many applications, including imaging systems and high-speed analog-to-digital converters, such a saving can be dramatic; however, the dimensionality reduction from N to M is still not on par with state-of-the-art transform coding systems. While many natural and manmade signals can be described to a first-order as sparse or compressible, their sparse supports often have an underlying domain specific structure [3–6]. Exploiting this structure in CS recovery has two immediate benefits. First, the number of compressive measurements required for stable recovery decreases due to the reduction in the degrees of freedom of a sparse or compressible signal. Second, true signal information can be better differentiated from recovery artifacts during signal recovery, which increases recovery robustness. Only by exploiting a priori information on coefficient structure in addition to signal sparsity, can CS hope to be competitive with the state-of-the-art transform coding algorithms for dimensionality reduction.

Fortunately, it is possible to design CS recovery algorithms that exploit the knowledge of structured sparsity models with provable performance guarantees [3, 5, 6]. In particular, the model-based CS recovery framework in [3] generalizes to any structured-sparsity model that has a tractable model-based approximation algorithm. This framework has been applied productively to two structured signal models: block sparsity and wavelet trees with robust recovery guarantees from $\mathcal{O}\left(K\right)$ measurements [3]. To recover signals that have structured sparsity, problem-specific convex relaxation approaches are also used in the literature with recovery guarantees similar to those in [3]; e.g., for block sparse signals, see [5, 6].

In this paper, we introduce a new structured sparsity model, called the (K, C)-model, that constrains the K-sparse signal coefficients to be contained within at most C-clusters. In contrast to the block sparsity model in [5, 6], our proposed model does not assume prior knowledge of the locations and sizes of the coefficient clusters. We show that $\mathcal{O}\left(K + C\log(N/C)\right))$ random projections are sufficient for (K, C)-model signal recovery using a subspace counting argument. We also provide a polynomial-time model-based approximation algorithm based on dynamic programming and a CS recovery algorithm based on the model-based recovery framework of [3]. In contrast to the clustered sparse recovery algorithm based on the probabilistic Ising model in [7], the (K, C)-model has provable performance guarantees.

The paper is organized as follows. Section 2 provides the necessary theoretical and algorithmic background on model-based CS. Section 3 introduces the (K, C)-model, derives its sampling bound for CS recovery, and describes a dynamic programming solution for optimal (K, C)-model approximation. Section 4 discusses the aspect of compressibility and highlights some connections to the block sparsity model. Simulation results are given in Section 5 to demonstrate the effectiveness of the (K, C)-model. Section 6 provides our conclusions.

2. Model-based CS Background

A K-sparse signal vector x lives in $\Sigma_K \subset \mathbb{R}^N$, which is a union of $\binom{N}{K}$ subspaces of dimension K. Other than its K-sparsity, there are no further constraints on the support or values of its coefficients. A *union-of-subspaces*

signal model (a *signal model* in the sequel for brevity) endows the K-sparse signal x with additional structure that allows certain K-dimensional subspaces in Σ_K and disallows others [4, 8].

More formally, let $x_{|\Omega}$ represent the entries of x corresponding to the set of indices $\Omega \subseteq \{1, \ldots, N\}$, and let Ω^C denote the complement of the set Ω. A signal model $\mathcal{M}_K$ is then defined as the union of m_K canonical K-dimensional subspaces

$$\mathcal{M}_K = \bigcup_{m=1}^{m_K} \mathcal{X}_m,\ \mathcal{X}_m := \{x : x_{|\Omega_m} \in \mathbb{R}^K, x|_{\Omega_m^C} = 0\}.$$

Each subspace $\mathcal{X}_m$ contains all signals x with $\mathrm{supp}(x) \in \Omega_m$. Thus, the signal model $\mathcal{M}_K$ is defined by the set of possible supports $\{\Omega_1, \ldots, \Omega_{m_K}\}$. Signals from $\mathcal{M}_K$ are called *K-model sparse*. Likewise, we may define $\mathcal{M}_K^c$ to be the set of c-wise differences of signals belonging to $\mathcal{M}_K$. Clearly, $\mathcal{M}_K \subseteq \Sigma_K$ and $\mathcal{M}_K^4 \subseteq \Sigma_{4K}$. In the sequel, we will use an algorithm $\mathbb{M}(x; K)$ that returns the best K-term approximation of the signal x under the model $\mathcal{M}_K$.

If we know that the signal x being acquired is K-model sparse, then we can relax the standard restricted isometry property (RIP) [1] of the CS measurement matrix Φ and still achieve stable recovery from the compressive measurements $y = \Phi x$. The *model-based* RIP $\mathcal{M}_K$-RIP requires that

$$(1 - \delta_{\mathcal{M}_K})\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta_{\mathcal{M}_K})\|x\|_2^2 \qquad (1)$$

hold for signals $x \in \mathcal{M}_K$ [4, 8], where $\delta_{\mathcal{M}_K}$ is the model-based RIP constant.

Blumensath and Davies [4] have quantified the number of measurements M necessary for a subgaussian CS matrix to have the $\mathcal{M}_K$-RIP with constant $\delta_{\mathcal{M}_K}$ and with probability $1 - e^{-t}$ to be

$$M \geq \frac{2}{c\delta^2_{\mathcal{M}_K}} \left(\ln(2m_K) + K \ln \frac{12}{\delta_{\mathcal{M}_K}} + t \right). \qquad (2)$$

This bound can be used to recover the conventional CS result by substituting $m_K = \binom{N}{K} \approx (Ne/K)^K$.

To take practical advantage of signal models in CS, we can integrate them into a standard CS recovery algorithm based on iterative greedy approximation. The key modification is surprisingly simple [3]: we merely replace the best K-term approximation step with the best K-term model-based approximation $\mathbb{M}(x; K)$. For example, in the CoSaMP algorithm [9], the best LK-term approximation (with L a small integer) is modified to incorporate a best LK-term model-based approximation. The resulting algorithm (see [3]) then inherits the following model-based CS recovery guarantee at each iteration i, when the measurement matrix Φ has the $\mathcal{M}_K^4$-RIP with $\delta_{\mathcal{M}_K^4} \leq 0.1$:

$$\|x - \widehat{x}_i\|_2 \leq 2^{-i}\|x\|_2 + 20\Bigg(\|x - x_{\mathcal{M}_K}\|_2 + \frac{1}{\sqrt{K}}\|x - x_{\mathcal{M}_K}\|_1 + \|n\|_2 \Bigg),$$

where $x_{\mathcal{M}_K} = \mathbb{M}(x; K)$ is the best model-based approximation of x within $\mathcal{M}_K$.

3. The (K, C)-Model

Motivation: The *block sparsity model* is used in applications where the significant coefficients of a sparse signal appear in designated blocks on the ambient signal dimension, e.g., group sparse regression problems, DNA microarrays, MIMO channel equalization, source localization in sensor networks, and magnetoencephalography [3, 5, 6, 10–14]. It has been shown that recovery algorithms provably improve standard CS recovery by exploiting this block-sparse structure [3, 5].

The (K, C)-model generalizes the block sparsity model by allowing the significant coefficients of a sparse signal to appear in at most C clusters of unknown size and location (Figure 1(a)). This way, the (K, C)-model further accommodates additional applications in, e.g., neuroscience problems that are involved with decoding of natural images in the primary visual cortex (V1) or understanding the statistical behavior of groups of neurons in the retina [15]. In this section, we formulate the (K, C)-model as a union of subspaces and pose an approximation algorithm on this union of subspaces.

To define the set of (K, C)-sparse signals, without loss of generality, we focus on canonically sparse signals in $N + 2$ dimensions whose first and last coefficients are zero. Consider expressing the support of such signals via run-length coding with a vector $\beta = (\beta_1, \ldots, \beta_{2C+1})$ ($\beta_j \neq 0$), where β_{odd} counts the number of continuous zero-signal values and β_{even} counts the number of continuous nonzero-signal values (i.e., clusters).

Definition: The (K, C)-sparse signal model $\mathcal{M}_{(K,C)}$ is defined as

$$\mathcal{M}_{(K,C)} = \left\{ x \in R^{N+2} \,\middle|\, \sum_{i=1}^{2C+1} \beta_i = N + 2, \sum_{i=1}^{C} \beta_{2i} = K \right\}. \qquad (3)$$

Sampling Bound: The number of subspaces $m_{(K,C)}$ in $\mathcal{M}_{(K,C)}$ can be obtained by counting the number of *positive* solutions to the following integer equations:

$$\begin{aligned} \beta_1 + \beta_2 + \ldots + \beta_{2C+1} &= N + 2, \\ \beta_2 + \beta_4 + \ldots + \beta_{2C} &= K, \end{aligned}$$

which can be rewritten as

$$\begin{aligned} \beta_1 + \beta_3 + \ldots + \beta_{2C+1} &= N + 2 - K, \\ \beta_2 + \beta_4 + \ldots + \beta_{2C} &= K. \end{aligned} \qquad (4)$$

Note that the number of positive integer solutions to the following problem:

$$\beta_1 + \beta_2 + \beta_3 + \ldots + \beta_n = N,$$

is given by $\binom{N-1}{n-1}$. Then, we can count the solutions to the two of decoupled problems in (4) and multiply the number of solutions to obtain $m_{(K,C)}$:

$$m_{(K,C)} = \binom{N + 1 - K}{C} \binom{K - 1}{C - 1}. \qquad (5)$$

Plugging (5) into (2), we obtain the sampling bound for $\mathcal{M}_{(K,C)}$:

$$M = \mathcal{O}\left(K + C\log\frac{N}{C}\right). \tag{6}$$

Note that the (K,C)-sampling bound (6) becomes the standard CS bound of $M = \mathcal{O}\left(K\log\frac{N}{K}\right)$ when $C \approx K$.

Model Approximation Algorithm: In this section we focus on designing an algorithm $\mathbb{M}(x;K,C)$ for finding the best (K,C)-model approximation to a given signal x. The algorithm uses the principle of dynamic programming [16]. For simplicity, we focus on the problem of finding the *cost* of the best (K,C)-clustered signal approximation in ℓ_2. This solution generalizes to the best (K,C)-clustered signal approximation in ℓ_p for $p \geq 1$. The actual sparsity pattern can be then recovered using standard back-tracing techniques; see [16] for the details.

The algorithm $\mathbb{M}(x;K,C)$ computes an array $\text{cost}[i,j,k,c]$, where $1 \leq i \leq j \leq N$, $0 \leq k \leq K$, and $0 \leq c \leq C$. At the end of the algorithm, each entry $\text{cost}[i,j,k,c]$ contains the smallest cost of approximating $x_{i:j}$, the signal vector restricted to the index set $[i,\ldots,j]$, using at most k non-zero entries that span at most c clusters. $\mathbb{M}(x;K,C)$ performs the following operations.

(*Initialization*) When either $c = 0$ or $k = 0$, the signal approximation costs can be computed directly, since $\text{cost}[i,j,0,c] = \|x_{i:j}\|_2^2$ and $\text{cost}[i,j,k,0] = \|x_{i:j}\|_2^2$, for all valid indices i,j,k,c. Moreover, for all entries i,j,k,c such that $c > 0$ and $j - i + 1 \leq k$, we have $\text{cost}[i,j,k,c] = 0$ since we can include all $j-i+1$ coordinates of the vector $x_{i:j}$ in the approximation.

(*Main loop*) All other cost entries can then be computed using the following recursion:

$$\text{cost}[i,j,k,c] = \min_{c^\star=0\ldots c}\ \min_{k^\star=0\ldots k}\ \min_{j^\star=i\ldots j-1} \left\{\text{cost}[i,j^\star,k^\star,c^\star] \times \text{cost}[j^\star+1,j,k-k^\star,c-c^\star]\right\}.$$

The correctness of the algorithm follows from the following observation. Let $\overline{v}$ be the best (k,c)-clustered approximation of $x_{i:j}$. Unless all entries of $x_{i:j}$ can be included in the approximation $\overline{v}$ (in which case $j-i+1 \geq k$ and the entry has been already computed during initialization), then there must exist an index $l \in [i,\ldots,j]$ such that x_l is not included in $\overline{v}$. Let $l^\star = l$ if $l < j$, and $l^\star = j-1$ otherwise. Let $k^\star$ be the number of non-zero entries present in the *left segment* of $\overline{v}_{i:l^\star}$, and let $c^\star$ be the number of clusters present in that left segment. Then, it must be the case that $\overline{v}_{i:l^\star}$ is the best $(k^\star,c^\star)$-approximation to $x_{i:l}$, and $\overline{v}_{l+1:j}$ is the best $(k-k^\star,c-c^\star)$-approximation to $x_{(l^\star+1):j}$. Otherwise, those better approximations could have been concatenated together to yield an even better (k,c)-approximation of $x_{i:j}$. Thus, the recursive formula will identify the optimal split and compute the optimal approximation cost.

The cost table contains $\mathcal{O}\left(N^2KC\right)$ entries. Each entry can be computed in $\mathcal{O}\left(NKC\right)$ time. Thus, the running time of the algorithm is $\mathcal{O}\left(N^3K^2C^2\right)$.

4. Additional Remarks

Compressibility: Just as compressible signals are *nearly K-sparse* and live close to the union of subspaces Σ_K in $\mathbb{R}^N$, (K,C)-compressible signals are *nearly (K,C)-model sparse* and live close to the restricted union of subspaces $\mathcal{M}_{(K,C)}$. Here, we rigorously introduce a (K,C)-compressible signal model in terms of the decay of their (K,C)-model approximation error.

We first define the ℓ_2 error incurred by approximating $x \in \mathbb{R}^N$ by the best approximation in $\mathcal{M}_{(K,C)}$:

$$\sigma_{\mathcal{M}_{(K,C)}}(x) \triangleq \inf_{\bar{x}\in\mathcal{M}_{(K,C)}} \|x-\bar{x}\|_2 = \|x-\mathbb{M}(x;K,C)\|_2.$$

The decay of the (K,C)-model approximation error in (7) defines the (K,C)-compressibility of a signal. Then, a set of (K,C)-*model s-compressible signals* is given by

$$\mathfrak{M}_s = \left\{ x \in \mathbb{R}^N : \sigma_{\mathcal{M}_{j(K,C)}}(x) \leq S(jK)^{-1/s}, \quad 1 \leq K \leq N, S < \infty, j = 1,\ldots,\left\lfloor\frac{N}{K}\right\rfloor \right\}. \tag{7}$$

Define $S_{\mathfrak{M}}$ as the smallest value of S for which this condition holds for x and s.

We use the restricted amplification property (RAmP) and the nested approximation property (NAP) in [3] to ensure that the (K,C)-model based CoSaMP recovery possesses the following guarantee for (K,C)-model s-compressible signals at each iteration i:

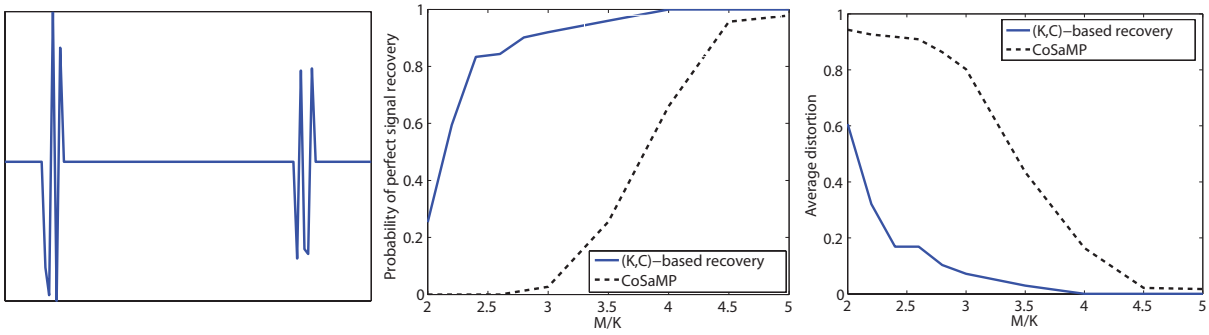
$$\|x-\widehat{x}_i\|_2 \leq 2^{-i}\|x\|_2 + 35\left(\|n\|_2 + \frac{S_{\mathfrak{M}}}{K^s}(1+\ln\lceil N/K\rceil)\right), \tag{8}$$

when Φ has the $\mathcal{M}^4_{(K,C)}$-RIP with $\delta_{\mathcal{M}^4_{(K,C)}} \leq 0.1$ and the (ϵ_K, r)-RAmP with $\epsilon_K \leq 0.1$ and $r = s-1$.

Simulation via Block Sparsity: It is possible to recover (K,C)-sparse signals by using the block sparsity model if we are willing to pay an added penalty in terms of the number of measurements. To demonstrate this, we define uniform blocks of size K/C (e.g., average cluster length) on the signal space. Then, it is straightforward to see that the number of active blocks B in the block sparse model is upper-bounded by

$$B \leq 2(C-1) + \frac{K-2(C-1)}{K/C} \leq 3C. \tag{9}$$

To reach this upper bound, we first construct a (K,C)-sparse signal that has $(C-1)$-clusters with 2 coefficients and a single cluster with the remaining sparse coefficients. We then place the clusters with two coefficients at the boundary of the block sparse model so that each cluster activate two blocks in the block sparse model to arrive at (9). Then, the (K,C)-equivalent block sparse model requires $M = \mathcal{O}\left(BK/C + B\log\frac{N}{B}\right)$ samples, where $B = 3C$.



(a) A $(10, 2)$-model signal (b) Reconstruction probability (c) Recovery error

Figure 1: *Monte Carlo simulation results for (K, C)-model based recovery with $K = 10, C = 2$..*

5. Experiments

In this section we demonstrate the performance of (K, C)-model based recovery. Our test signals are the class of length-100 clustered-sparse signals with $K = 10, C = 2$. We run both the CoSaMP algorithm as well as (K, C)-model based CoSaMP algorithm [3] until convergence for 1000 independent trials. In Fig. 1(a), a sample realization of the signal is displayed. It is evident from Figs. 1(b) and (c) that enforcing the structured sparsity model in the recovery process significantly improves CS reconstruction performance. In particular, Fig. 1(b) demonstrates that approximately 85% of the signals are almost perfectly recovered at $M = 2.5K$, whereas CoSaMP fails to recover any signals at this level of measurements. Instead, traditional sparsity-based recovery requires $M \geq 4.5K$ to attain comparable performance. Similarly, Figure. 1(c) displays the rapid decrease in average recovery distortion of our proposed method, as compared to the conventional approach. The (K, C)-sparse approximation algorithm codes are available at dsp.rice.edu/software/KC.

6. Conclusions

In this paper, we have introduced a new sparse signal model that generalizes the block-sparsity model used in the CS literature. To exploit the provable model-based CS recovery framework of [3], we developed a dynamic programming algorithm that computes, for any given signal, its optimal ℓ_2-approximation within our clustered sparsity model. We then demonstrated that significant performance gains can be made by exploiting the clustered signal model beyond the simplistic sparse model that are prevalent the CS literature.

Acknowledgments

The authors would like to thank Marco F. Duarte for useful discussions and Andrew E. Waters for converting the (K, C)-model MATLAB code into C++. VC, CH and RGB were supported by the grants NSF CCF-0431150 and CCF-0728867, DARPA/ONR N66001-08-1-2065, ONR N00014-07-1-0936 and N00014-08-1-1112, AFOSR FA9550-07-1-0301, ARO MURI W311NF-07-1-0185, and the Texas Instruments Leadership University Program. PI is supported in part by David and Lucille Packard Fellowship and by MADALGO (Center for Massive Data Algorithmics, funded by the Danish National Research Association) and by NSF grant CCF-0728645.

References:

[1] E. J. Candès, "Compressive sampling," in *Proc. International Congress of Mathematicians*, vol. 3, (Madrid, Spain), pp. 1433–1452, 2006.

[2] D. L. Donoho, "Compressed sensing," *IEEE Trans. Info. Theory*, vol. 52, pp. 1289–1306, Sept. 2006.

[3] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," 2008. Preprint. Available at http://dsp.rice.edu/cs.

[4] T. Blumensath and M. E. Davies, "Sampling theorems for signals from the union of finite-dimensional linear subspaces," *IEEE Trans. Info. Theory*, Dec. 2008.

[5] Y. Eldar and M. Mishali, "Robust recovery of signals from a union of subspaces," 2008. Preprint.

[6] M. Stojnic, F. Parvaresh, and B. Hassibi, "On the reconstruction of block-sparse signals with an optimal number of measurements," Mar. 2008. Preprint.

[7] V. Cevher, M. F. Duarte, C. Hegde, and R. G. Baraniuk, "Sparse signal recovery using Markov Random Fields," in *Proc. Workshop on Neural Info. Proc. Sys. (NIPS)*, (Vancouver, Canada), Dec. 2008.

[8] Y. M. Lu and M. N. Do, "Sampling signals from a union of subspaces," *IEEE Signal Processing Mag.*, vol. 25, pp. 41–47, Mar. 2008.

[9] D. Needell and J. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, June 2008.

[10] J. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Processing*, vol. 86, pp. 572–588, Apr. 2006.

[11] Y. Kim, J. Kim, and Y. Kim, "Blockwise sparse regression," *Statistica Sinica*, vol. 16, no. 2, p. 375, 2006.

[12] L. Meier, S. van de Geer, and P. Buhlmann, "The group lasso for logistic regression," *Journal of Royal Stat. Society: Series B (Statistical Methodology)*, vol. 70, no. 1, pp. 53–71, 2008.

[13] F. Parvaresh, H. Vikalo, S. Misra, and B. Hassibi, "Recovering Sparse Signals Using Sparse Measurement Matrices in Compressed DNA Microarrays," *IEEE Journal of Selected Topics in Sig. Proc.*, vol. 2, no. 3, pp. 275–285, 2008.

[14] I. F. Gorodnitsky, J. S. George, and B. D. Rao, "Neuromagnetic source imaging with FOCUSS: a recursive weighted minimum norm algorithm," *Electroenceph. and Clin. Neurophys.*, vol. 95, no. 4, pp. 231–251, 1995.

[15] P. J. Garrigues and B. A. Olshausen, "Learning Horizontal Connections in a Sparse Coding Model of Natural Images," in *Advances in Neural Info. Proc. Sys. (NIPS)*, 2008.

[16] T. H. Corman, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*. MIT Press and McGraw-Hill, New York, USA, 2001.

Special session on

Compressed Sensing

Chair: Yonina ELDAR

Compressed sensing signal models - to infinity and beyond?

Thomas Blumensath and Michael Davies

IDCOM & Joint Research Institute for Signal and Image Processing
Edinburgh University, King's Buildings, Mayfield Road, Edinburgh, UK
thomas.blumensath@ed.ac.uk, mike.davies@ed.ac.uk

Abstract:

Compressed sensing is an emerging signal acquisition technique that enables signals to be sampled well below the Nyquist rate, given a finite dimensional signal with a sparse representation in some orthonormal basis. In fact, sparsity in an orthonormal basis is only one possible signal model that allows for sampling strategies below the Nyquist rate. We discuss some recent results for more general signal models based on unions of subspaces that allow us to consider more general *structured representations*. These include classical sparse signal models and finite rate of innovation systems as special cases.
We consider the dimensionality conditions for two aspects of the compressed sensing inverse problem: the existence of *one-to-one* maps to lower dimensional observation spaces and the smoothness of the inverse map.
On the surface Lipschitz smoothness of the inverse map appears to limit the applicability of compressed sensing to infinite dimensional signal models. We therefore discuss conditions where smooth inverse maps are possible even in infinite dimensions. Finally we conclude by mentioning some recent work [14] which develops the these ideas further allowing the theory to be extended beyond exact representations to *structured approximations*.

1. Introduction

Since Nyquist and Shannon we are used to sampling continuous signals at a rate that is twice the bandwidth of the signal. However recently, under the umbrella title of *compressed sensing*, researchers have begun to explore how and when signals can be recovered using much fewer samples, but relying on known signal structure. Importantly the papers by Candes, Romberg and Tao [4], [5], [6] and by Donoho [8] have shown that under certain conditions on the signal sparsity and the sampling operator (which are often satisfied by certain random matrices), finite dimensional signals can be stably reconstructed when the number of observations is of the order of the signal sparsity and only logarithmically dependent on the ambient space dimension. Furthermore the reconstruction can be performed using practical polynomial time algorithms.
Here we discuss a generalization of the sparse signal model that enables us to consider more structured signal types. We are interested in when the signals can be stably reconstructed (or in some cases approximated). We finish the paper by considering the implications of these results for ∞-dimensional signal models and extending from *structured representations* to *structured approximation*.

2. Signal models and problem statement

The problem can be formulated as follows. A continuous or discrete signal f from some separable Hilbert space is to be sampled. This is done by using M linear measurements $\{\langle f, \phi_n \rangle\}_n$, where $\langle \cdot, \cdot \rangle$ is the inner product and where $\{\phi_n\}$ is a set of vectors from the Hilbert space under consideration. Through the choice of an appropriate orthonormal basis, ψ we can replace f by the vector x such that $f = \sum_{i=1}^{N} \psi_i x_i$. Let $\Phi \in \mathbb{R}^{M \times N}$ be the sensing matrix with entries $\langle \psi_i, \phi_j \rangle$. The observation can then be written as

$$\mathbf{y} = \Phi \mathbf{x}. \tag{1}$$

In compressed sensing it is paramount to consider signals $\mathbf{x}$ that are highly structured and in the original papers, $\mathbf{x}$ was assumed to be an *exact k-sparse* vector, i.e. a vector with not more than k non-zero entries (we discuss a relaxation of this in section 6.). This naturally defines the signal model as a union of N-choose-k k-dimensional subspaces, $\mathcal{K}$.
A nice generalization of this model, introduced in [12], is to consider the signal $\mathbf{x}$ to be an element from a union of arbitrary subspaces $\mathcal{A}$, defined formally as

$$\mathcal{A} = \bigcup_{j}^{L} S_j, \; S_j = \{\mathbf{y} = \Omega_j \mathbf{a}, \Omega_j \in \mathbb{R}^{N \times k_j}, \mathbf{a} \in \mathbb{R}^{k_j}\}, \tag{2}$$

where the Ω_j are bases for linear subspaces. This general signal model incorporates many previously considered compressed sensing settings, including:

- The exact k-sparse signal model, $\mathcal{K}$
- Finite Rate of Innovation (FRI) [15] signal models, if we allow an uncountable number of subspaces (e.g. filtered streams of Dirac functions)
- signals that are k-sparse in a general, possibly redundant dictionary
- exact k-sparse signals whose non-zero elements form a tree

- multi-dimensional signals that are k-sparse with common support

Importantly this model allows us to incorporate additional structure which can in turn be advantageous by for example reducing signal complexity (as in the tree-constrained sparse model).
The aim of compressed sensing is to select a linear sampling operator, Φ, such that there exists a unique inverse map ${\Phi|_{\mathcal{K}}}^{-1} : \Phi(\mathcal{K}) \mapsto \mathcal{K}$. Moreover, for stability, we generally desire $\Phi(\mathcal{K})$ to be a bi-Lipschitz embedding of $\mathcal{K}$. In standard compressed sensing this stability is captured by the restricted isometry property [1].
When considering the union of subspaces model we can similarly look for a Φ with a unique stable (Lipschitz) inverse map ${\Phi|_{\mathcal{A}}}^{-1} : \Phi(\mathcal{A}) \mapsto \mathcal{A}$. Below we will discuss both necessary and sufficient conditions for this.

3. Existence of a unique inverse map

In [12] it was shown that a necessary condition for a unique inverse map to exist is that $M \geq M_{\min} := \max_{i \neq j} k_i + k_j$. If this is not the case we can find a vector $\mathbf{x} \in S_i \oplus S_j, \mathbf{x} \neq 0$ such that $\Phi x = 0$. The authors further go on to show that when there are a *countable* number of finite dimensional subspaces then the set of such sampling operators, Φ giving a unique inverse is dense.
In [3] we presented a slight refinement of this result for the case where the number of subspaces is finite. In this case *almost every* sampling operator, Φ, $M \geq M_{\min}$ has a unique inverse on $\mathcal{A}$. Furthermore even when $\max_i k_i < M < M_{\min}$ for almost every Φ the set of points in $\mathcal{A}$ without a unique inverse has zero measure (with respect to the largest subspace).
All this suggests that we might be able to perform compressed sensing from only slightly more observations than the dimension of the signal model, i.e. $M > \dim(\mathcal{A})$. Unfortunately we have so far ignored the issue of stability which we will see presents additional complications.

4. Stability of the inverse map

We now consider when the inverse mapping for the union of subspaces model is stable. Here we are particularly interested in the Lipschitz property of this inverse map and we derive conditions for the existence of a bi-Lipschitz embedding from $\mathcal{A}$ into a subset of $\mathbb{R}^M$.
The Lipschitz property is an important aspect of the map which ensures stability of any reconstruction to perturbations of the observation and in effect specifies the robustness of compressed sensing against noise and quantization errors. Furthermore, in the k-sparse model, the bi-Lipschitz property has also played an important role in demonstrating the existence of efficient and robust reconstruction algorithms through the k-restricted isometry property (RIP) [4, 5, 6, 8].
A natural extension of the k-restricted isometry for the union of subspaces model is [12, 3]:

Definition: ($\mathcal{A}$-restricted isometry) For any matrix Φ and any subset $\mathcal{A} \subset \mathbb{R}^N$ we define the $\mathcal{A}$-restricted isometry constant $\delta_{\mathcal{A}}(\Phi)$ to be the smallest quantity such that

$$(1 - \delta_{\mathcal{A}}(\Phi)) \leq \frac{\|\Phi \mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2} \leq (1 + \delta_{\mathcal{A}}(\Phi)), \qquad (3)$$

holds for all $\mathbf{x} \in \mathcal{A}$.

If we define the set $\bar{\mathcal{A}} = \{\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2 : \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{A}\}$ then $\delta_{\bar{\mathcal{A}}} < 1$ controls the Lipschitz constants of Φ and ${\Phi|_{\mathcal{A}}}^{-1}$ (in the standard compressed sensing this is directly equivalent to δ_{2m}). Specifically let us define:

$$\|\Phi(\mathbf{y}_1) - \Phi(\mathbf{y}_2)\|_2 \leq K_F \|\mathbf{y}_1 - \mathbf{y}_2\|_2 \qquad (4)$$
$$\|\Phi|_{\mathcal{A}}^{-1}(\mathbf{x}_1) - \Phi|_{\mathcal{A}}^{-1}(\mathbf{x}_2)\|_2 \leq K_I \|\mathbf{x}_1 - \mathbf{x}_2\|_2 \qquad (5)$$

then a straight forward consequence of the $\bar{\mathcal{A}}$-RIP definition is that:

$$K_F \leq \sqrt{1 + \delta_{\bar{\mathcal{A}}}} \qquad (6)$$
$$K_I \leq \frac{1}{\sqrt{1 - \delta_{\bar{\mathcal{A}}}}} \qquad (7)$$

Note, as always with RIP, it is prudent to consider appropriate scaling of Φ to balance the upper and lower inequalities in (3).
The following results, proved in [3], give neccessary and sufficient conditions for Φ to be an $\mathcal{A}$-restricted isometry.

4.1 Sufficient conditions

Theorem 1 *For any $t > 0$, let*

$$M \geq \frac{2}{c\delta_{\mathcal{A}}} \left(\ln(2L) + k \ln\left(\frac{12}{\delta_{\mathcal{A}}}\right) + t \right), \qquad (8)$$

then there exist a matrix $\Phi \in \mathbb{R}^{M \times N}$ and a constant $c > 0$ such

$$(1 - \delta_{\mathcal{A}}(\Phi))\|\mathbf{x}\|_2^2 \leq \|\Phi \mathbf{x}\|_2^2 \leq (1 + \delta_{\mathcal{A}}(\Phi))\|\mathbf{x}\|_2^2 \qquad (9)$$

holds for all $\mathbf{x}$ from the union of L arbitrary k dimensional subspaces $\mathcal{A}$. What is more, if Φ is generated by randomly drawing i.i.d. entries from an appropriately scaled sub-gaussian distribution then this matrix satisfies equation (9) with probability at least

$$1 - e^{-t}. \qquad (10)$$

The proof follows the same lines as the construction of random matrices with k-RIP [1].
In contrast to the previous results on the existence of a unique inverse map this sufficient condition is logarithmic in the number of subspaces considered.

4.2 Necessary conditions

We next show that the logarithmic dependence on L is in fact necessary. This can be done by considering the distance between the optimally packed unit norm vectors in $\mathcal{A}$ as a function of the number of observations. To this end it is useful to define a measure of separation between vectors in the different subspaces:

Definition: ($\Delta(\mathcal{A})$ subspace separation) Let $\mathcal{A} = \bigcup_i S_i$ be the union of subspaces S_i and let $\mathcal{A}/S_i$ be the union of subspaces with the i^{th} subspace excluded. The subspace separation of $\mathcal{A}$ is defined as

$$\Delta(\mathcal{A}) = \inf_i \left[\sup_{\substack{\mathbf{x}_i \in S_i \\ \|\mathbf{x}_i\|_2=1}} \left[\inf_{\substack{\mathbf{x}_j \in \mathcal{A}/S_i \\ \|\mathbf{x}_j\|_2=1}} \|\mathbf{x}_i - \mathbf{x}_j\|_2 \right] \right] \quad (11)$$

We can now state the following necessary condition for the existence of an $\mathcal{A}$-restricted isometry in terms of $\Delta(\mathcal{A})$ and the observation dimension.

Theorem 2 *Let $\mathcal{A}$ be the union of L subspaces of dimension no more than k. In order for a linear map $\Phi : \mathcal{A} \mapsto \mathbb{R}^N$ to exist such that it has a Lipschitz constant K_F and such that its inverse map ${\Phi_{\mathcal{A}}}^{-1} : \Phi(\mathcal{A}) \mapsto \mathcal{A}$ has a Lipschitz constant K_I, it is necessary that*

$$M \geq \frac{\ln(L)}{\ln\left(\frac{4K_F K_I}{\Delta(\mathcal{A})}\right)}. \quad (12)$$

Therefore, for a fixed subspace separation, the *necessary* number of samples grows logarithmically with the number of subspaces.

This last fact suggests that extending the compressed sensing framwork to infinite dimensional signals may be problematic. For example, it implies that the $\log(N)$ dependence in the standard k-sparse signal model is necessary (from the easily derived bound $\Delta(\mathcal{A}) \geq \sqrt{2/k}$) and therefore such a framework does not directly map to infinite dimensional signal models.

5. 2 routes to infinity

Most of the results in compressed sensing assume that the ambient signal space, N, is finite dimensional. This also implies in the case of the k-sparse signal model ($k < \infty$) that the number of subspaces, L, in the signal model is also finite. In fact we would ideally like to understand when we can perform compressed sensing when either or both the quantities, N and L, are infinite. Specifically when might a stable unique inverse for $\Phi|_{\mathcal{A}}$ exist based upon a finite number of observations.

For example the Finite Rate of Innovation (FRI) sampling framework introduced by Vetterli *et al.* [15] provides sampling strategies for signals composed of the weighted sum of a finite stream of diracs. In this case both N and L are uncountably infinite while $M > 2k$ is sufficient to reconstruct the signal.

Below we consider two possible routes to infinity and comment on their stability. Note other routes to infinity also exist, such as when we let k, M and $N \to \infty$ while keeping k/M and M/N finite [9], or in the blind multi-band signal model [10, 13], where the sampling rate, M/N, is finite but where $M, N \to \infty$.

5.1 k, L finite and N infinite

We begin with the easy case that the reader might consider to be a bit of a cheat.

Consider a signal model $\mathcal{A} \subset \mathcal{H}$, where $\mathcal{H}$ is an infinite dimensional separable Hilbert space (i.e. $N = \infty$). Assume that both k and L are finite. In this case the union of subspace model $\mathcal{A}$ automatically lives within a finite dimensional subspace, $U \subset \mathcal{H}$ defined as:

$$U := \bigoplus_{i=1}^{L} S_i \quad (13)$$

Note that $\dim(U) \leq kL < \infty$. We can therefore first project onto the finite dimensional subspace U and then apply the above theory to guarantee both the existence and stability of inverse mappings in this setting.

Two signal models that naturally fit into this framework are: the block-based sparsity model [11], which is related to the multiple measurement vectors problem and has been used recently in a blind multi-band signal acquisition scheme [13]; and the tree-based sparsity model where the usual k-sparse model is constrained to form a rooted subtree where $L \leq \frac{(2e)^k}{k+1}$ independent of N [3] and naturally occurs in multi-resolution modelling. This model has also been recently extended to include tree-compressible signals [14]: see section 6..

5.2 k finite, L and N infinite

From Theorem 2 the only way in which the number of subspaces can be infinite (or even un-countable) while permitting a stable inverse mapping, $\Phi|_{\mathcal{A}}^{-1}$, with M finite is if the subspace separation, $\Delta(\mathcal{A}) = 0$. In such a case the union of subspace model may often form a nonlinear signal manifold. Note also that when we have an uncountable union of k-dimensional subspaces the dimension of the signal model may well be greater than k.

As an example let us consider the case of a simple Finite Rate of Innovation process [15]. Such models can be described as an uncountable union of subspaces and the key existence results from [12] immediately apply. However this tells us nothing about stability. For simplicity we will limit ourselves to a basic form of periodic FRI signal on $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ which can be written as:

$$x(t) = G(\tau, \mathbf{a})(t) := \sum_{i=0}^{k-1} a_i \psi(t - \tau_i) \quad (14)$$

where ψ are also periodic on $\mathbb{T}$, $\tau = \{\tau_1, \ldots, \tau_k\}$ and $\mathbf{a} = \{a_1, \ldots, a_k\} \in \mathbb{R}^k$.

In [15] the possibility of a periodic Dirac stream is considered, i.e. $\psi(t) = \delta(t)$, $t \in [0, 1]$. Here we avoid the Dirac stream by restricting to the case where $\psi(t) \in \mathbf{L^2}(\mathbb{T})$ and directly consider the signal model defined by the parametric mapping:

$$G : U \times \mathbb{R}^k \mapsto \mathbf{L^2}(\mathbb{T}) \quad (15)$$

where $U = \{\tau \in \mathbb{R}^k : \tau_i < \tau_j, \forall i < j\}$. Individual subspaces can be identified with a given τ. Furthermore the continuity of the shift operator implies that for any $\psi(t) \in \mathbf{L^2}(\mathbb{T})$, the associated union of subspace model, $\mathcal{A}$ has $\Delta(\mathcal{A}) = 0$. Equivalently we can only find a finite number of subspaces, S'_j, whose union, $\mathcal{A}' := \bigcup_j^{L'} S'_j \subset$

$\mathcal{A}$ has $\Delta(\mathcal{A}') \geq \epsilon > 0$). Theorem 2 can then be used to lower bound the Lipschitz constants of any embedding in terms of the number of subspaces, L' of any such $\mathcal{A}'$.
We have seen that Theorem 2 does not preclude a stable embedding for such systems. However there is clearly more work needed to determine when such models can have finite dimensional stable embeddings. One possible avenue of research would be to examine the recently derived sufficient conditions for stable embedding of general smooth manifolds [7, 2].

6. ...and beyond?

In reality all the union of subspace models we have considered are an idealization. In practise we can expect to, at most, be able to *approximate* a signal by one from a union of subspaces model. In traditional compressed sensing this is the difference between finding a sparse representaion of an exact k-sparse signal and finding a good sparse approximation of a compressible signal (i.e. one that is well approximated by a k-sparse signal).
Recent work at Rice university [14] has shown that for the special case of restricted k-sparse models (such as the tree-restricted sparsity) the exact union of subspace model can be extended to approximate union of subspace models that are subsets of compressible signal models.
In order to go beyond exact representations further conditions are introduced. Notably:

1. *Nested Approximation Property (NAP)* - this specifies sets of models, $\mathcal{M}_K$, that are naturally nested.

2. *Restricted Amplification Property (RAmP)* - this imposes additional regularity on the sensing matrix Φ when acting on the difference between the $\mathcal{M}_K$ subspaces and the $\mathcal{M}_{K-1}$ subspaces (in the k-sparse case it is interesting to note that the RAmP condition is automatically satisfied by the k-RIP condition).

There are therefore a number of interesting open questions. For example, are such additional conditions typically necessary to go beyond exact subspace representations? Furthermore can these additional tools be applied successfully to arbitrary union of subspace models (i.e. ones that are not subsets of the standard k-sparse model)?

7. Acknowledgements

This research was supported by EPSRC grants D000246/1 and D002184/1. MED acknowledges support of his position from the Scottish Funding Council and their support of the Joint Research Institute with the Heriot-Watt University as a component part of the Edinburgh Research Partnership.

References:

[1] R. Baraniuk, M. Davenport, R. De Vore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

[2] R Baraniuk and M Wakin. Random projections of smooth manifolds. *Foundations of Computational Mathematics*, 2007.

[3] T. Blumensath and M. E. Davies. Sampling theorems for signals from the union of linear subspaces. *Awaiting Publication, IEEE Transactions on Information Theory*, 2008.

[4] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, Feb 2006.

[5] Emmanuel Candès and Justin Romberg. Quantitative robust uncertainty principles and optimally sparse decompositions. *Foundations of Comput. Math*, 6(2):227 – 254, 2006.

[6] Emmanuel Candès and Terence Tao. Near optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. on Information Theory*, 52(12):5406 – 5425, 2006.

[7] K. L. Clarkson. Tighter bounds for random projections of manifolds. In *Proceedings of the twenty-fourth annual symposium on Computational geometry*, pages 39–48, 2008.

[8] D. Donoho. Compressed sensing. *IEEE Trans. on Information Theory*, 52(4):1289–1306, 2006.

[9] D. Donoho and J. Tanner. Counting faces of randomly-projected polytopes when the projection radically lowers dimension. *Journal of the AMS*, 2009.

[10] Y. C. Eldar. Compressed sensing of analog signals. *submitted to IEEE Trans. on Signal Processing*, 2008.

[11] Y. C. Eldar and M. Mishali. Robust recovery of signals from a union of subspaces. *Submitted to IEEE Trans. Inf Theory*, arXiv.org 0807.4581, 2008.

[12] Y. Lu and M. Do. A theory for sampling signals from a union of subspaces. *IEEE transactions on signal processing*, 56(6):2334–2345, 2008.

[13] M. Mishali and Y. C. Eldar. Blind multiband signal reconstruction: Compressed sensing for analog signals. *IEEE Trans. Signal Proc.*, 57(3):993–1009, 2009.

[14] M.F. Duarte R. G. Baraniuk, V. Cevher and C. Hegde. Model based compressed sensing. *Submitted to IEEE Transactions on Information Theory*, 2008.

[15] M. Vetterli, P. Marziliano, and T. Blu. Sampling signals with finite rate of innovation. *IEEE Transactions on Signal Processing*, 50(6):1417–1428, 2002.

Special session on

Frame Theory and Oversampling

Chair: Bernhard BODMANN

Gradient descent of the frame potential

Peter G. Casazza [1] and Matthew Fickus [2]

(1) Department of Mathematics, University of Missouri, Columbia, MO 65211 USA.
(2) Department of Mathematics & Statistics, Air Force Institute of Technology, WPAFB, OH 45433 USA.
pete@math.missouri.edu, matthew.fickus@afit.edu

Abstract:

Unit norm tight frames provide Parseval-like decompositions of vectors in terms of possibly nonorthogonal collections of unit norm vectors. One way to prove the existence of unit norm tight frames is to characterize them as the minimizers of a particular energy functional, dubbed the frame potential. We consider this minimization problem from a numerical perspective. In particular, we discuss how by descending the gradient of the frame potential, one, under certain conditions, is guaranteed to produce a sequence of unit norm frames which converge to a unit norm tight frame at a geometric rate. This makes the gradient descent of the frame potential a viable method for numerically constructing unit norm tight frames.

1. Introduction

The *analysis* operator of some finite sequence of vectors $\{f_m\}_{m=1}^M$ in an N-dimensional Hilbert space $\mathbb{H}_N$ is the operator $F : \mathbb{H}_N \to \mathbb{C}^M$, $(Ff)(m) := \langle f, f_m\rangle$. The corresponding *frame* operator is $F^*F : \mathbb{H}_N \to \mathbb{H}_N$,

$$F^*Ff = \sum_{m=1}^M \langle f, f_m\rangle f_m.$$

Generally speaking, *frame theory* is the study of how $\{f_m\}_{m=1}^M$ may be chosen in order to guarantee that F^*F is well-conditioned. In particular, $\{f_m\}_{m=1}^M$ is a *frame* for $\mathbb{H}_N$ if there exists *frame bounds* $0 < A \leq B < \infty$ such that $A\mathrm{I} \leq F^*F \leq B\mathrm{I}$, and is a *tight frame* if $A = B$, that is, if $F^*F = A\mathrm{I}$.

Typically, one's choice of f_m's is restricted according to some nonlinear, application-specific constraints. Of particular interest is the case of *unit norm tight frames*, that is, tight frames for which $\|f_m\| = 1$ for all $m = 1, \ldots, M$; such frames, known to exist for any $M \geq N$, provide Parseval-like decompositions in terms of vectors of unit length, even though these vectors are possibly nonorthogonal. Despite an ever-growing list of specific constructions of such frames, little is known about the manifold structure of the set of all unit norm tight frames.

In the hunt for unit norm tight frames, the *frame potential*, specifically defined as:

$$\mathrm{FP}(\{f_m\}_{m=1}^M) := \sum_{m,m'=1}^M |\langle f_m, f_{m'}\rangle|^2$$

for any sequence $\{f_m\}_{m=1}^M \in \mathbb{H}_N^M$, is a useful tool. Specifically, the frame potential quantifies the *total orthogonality* of a system of vectors by measuring the total potential energy stored within that system under a certain force which encourages orthogonality. Regarded as a functional over

$$\mathbb{S}_N^M = \big\{\{f_m\}_{m=1}^M \in \mathbb{H}_N^M : \|f_m\| = 1,\ m = 1, \ldots, M\big\},$$

one may show that when $M \geq N$, the local minimizers of the frame potential are precisely the unit norm tight frames of M elements for $\mathbb{H}_N$. In particular, as the frame potential is continuous and $\mathbb{S}_N^M$ is compact, one may conclude that such frames indeed exist for any $M \geq N$.

In this paper, we consider the minimization of the frame potential from a numerical perspective. In particular, in the next section, we compute the gradient of the frame potential, namely a specific direction $\{g_m\}_{m=1}^M \in \mathbb{H}_N^M$ in which to push $\{f_m\}_{m=1}^M$ so as to achieve the greatest instantaneous decrease of FP. Then, in an improvement over typical uses of gradient descent, we compute an exact step size in which to travel in this direction so as to produce a certain decrease in potential. In the third section, we estimate the size of this decrease in relation to how far the frame potential is from its minimum; under sufficient conditions, this estimate may be used to show that by descending the gradient of the frame potential, one may produce a sequence of unit norm frames which converge to a unit norm tight frame at a geometric rate.

The frame potential was introduced in [1], with its domain of optimization being later generalized in [4]. It has been used to characterize tight filter bank frames [5, 6]. The frame potential may also be used to prove the existence of tight fusion frames [3], and the local minimizers of the fusion frame potential are themselves a subject of interest [7, 9]. Further generalizations of the frame potential are considered in [2, 8].

2. The gradient of the frame potential

Our goal is to numerically minimize the frame potential over $\mathbb{S}_N^M$. As our domain of optimization is a product of spheres as opposed to the entire space $\mathbb{H}_N^M$, our approach departs from the classical theory of gradients. In particular, given $\{f_m\}_{m=1}^M \in \mathbb{S}_N^M$ and any $\{g_m\}_{m=1}^M \in \mathbb{H}_N^M$ such that $\langle f_m, g_m\rangle = 0$ for all $m = 1, \ldots, M$, we shall compute the rate of change of the frame potential as each f_m

is pushed along a great circle with tangent velocity g_m. We then define the gradient of FP to be that particular $\{g_m\}_{m=1}^M$ which makes this directional derivative as large as possible. We begin with the following result, which gives the first two derivatives of the frame potential of a single parameter family of frames:

Lemma 1 (Lemma 2 of [3])**.** *For any set of twice-differentiable parameterized curves $\{f_m(\cdot)\}_{m=1}^M$ in $\mathbb{H}_N$, the first two derivatives of $\varphi(t) := \mathrm{FP}(\{f_m(t)\}_{m=1}^M)$ are:*

$$\begin{aligned}\dot{\varphi}(t) &= 4\mathrm{ReTr}\big(\dot{F}(t)F^*(t)F(t)F^*(t)\big)\big),\\ \ddot{\varphi}(t) &= 4\mathrm{ReTr}\big(\ddot{F}(t)F^*(t)F(t)F^*(t)\big)\\ &\quad + 4\|\dot{F}(t)F^*(t)\|_{\mathrm{HS}}^2\\ &\quad + 2\|\dot{F}^*(t)F(t) + F^*(t)\dot{F}(t)\|_{\mathrm{HS}}^2,\end{aligned}$$

where $\dot{F}(t)$ and $\ddot{F}(t)$ are the analysis operators of $\{\dot{f}_m(t)\}_{m=1}^M$ and $\{\ddot{f}_m(t)\}_{m=1}^M$, respectively.

We now use Lemma 1 along with Taylor's theorem to asymptotically estimate the change in frame potential one obtains by perturbing a given $\{f_m\}_{m=1}^M \in \mathbb{S}_N^M$ along any choice of great circles. To be precise, letting:

$$\oplus f_m^\perp := \big\{\{g_m\}_{m=1}^M \in \mathbb{H}_N^M : \langle f_m, g_m\rangle = 0,\ \forall m\big\},$$

we have the following:

Theorem 2. *For any $\{f_m\}_{m=1}^M \in \mathbb{S}_N^M$, $\{g_m\}_{m=1}^M \in \oplus f_m^\perp$, let:*

$$f_m(t) := \cos(\|g_m\|t)f_m + (\sin(\|g_m\|t)/\|g_m\|)\, g_m$$

whenever $g_m \neq 0$ and let $f_m(t) := f_m$ otherwise. Then, $\{f_m(t)\}_{m=1}^M \in \mathbb{S}_N^M$ for any $t \in \mathbb{R}$, and satsifies:

$$\sum_{m=1}^M \|f_m(t) - f_m\|^2 \leq t^2 \sum_{m=1}^M \|g_m\|^2, \tag{1}$$

as well as:

$$\begin{aligned}\mathrm{FP}(\{f_m(t)\}_{m=1}^M) &\leq \mathrm{FP}(\{f_m\}_{m=1}^M)\\ &\quad + 4t\mathrm{Re}\sum_{m=1}^M \langle F^*Ff_m, g_m\rangle\\ &\quad + 8Mt^2\sum_{m=1}^M \|g_m\|^2.\end{aligned} \tag{2}$$

Proof. It is straightforward to show that $\|f_m(t)\| = 1$ for all $m = 1, \ldots, M$ and all $t \in \mathbb{R}$. To show (1), note that for any m such that $g_m \neq 0$, we have:

$$\begin{aligned}\|f_m(t) - f_m\|^2 &= \big(\cos(\|g_m\|t) - 1\big)^2 + \sin^2(\|g_m\|t)\\ &= 4\sin^2(\|g_m\|t/2)\\ &\leq \|g_m\|^2t^2.\end{aligned} \tag{3}$$

As (3) also immediately holds for any m such that $g_m = 0$, we may sum (3) over all m to conclude (1). To show (2), we apply Taylor's theorem to $\varphi(t) = \mathrm{FP}(\{f_m(t)\}_{m=1}^M)$ at $t = 0$:

$$\varphi(t) \leq \varphi(0) + t\dot{\varphi}(0) + \tfrac{1}{2}t^2 \max_{s\in\mathbb{R}} |\ddot{\varphi}(s)|. \tag{4}$$

To compute the terms in (4), note that

$$\dot{f}_m(t) = -\|g_m\|\sin(\|g_m\|t)f_m + \cos(\|g_m\|t)g_m \tag{5}$$

for any m such that $g_m \neq 0$. As (5) also immediately holds when $g_m = 0$, we have $\dot{f}_m(0) = g_m$ for all m. Thus, by Lemma 1,

$$\begin{aligned}\dot{\varphi}(0) &= 4\mathrm{ReTr}\big(\dot{F}(0)F^*(0)F(0)F^*(0)\big)\\ &= 4\mathrm{ReTr}\big(\dot{F}(0)F^*FF^*\big)\\ &= 4\mathrm{Re}\sum_{m=1}^M \langle \dot{F}(0)F^*FF^*e_m, e_m\rangle\\ &= 4\mathrm{Re}\sum_{m=1}^M \langle F^*Ff_m, \dot{f}_m(0)\rangle\\ &= 4\mathrm{Re}\sum_{m=1}^M \langle F^*Ff_m, g_m\rangle.\end{aligned} \tag{6}$$

Next, as taking the derivative of (5) yields $\ddot{f}_m(t) = -\|g_m\|^2 f_m(t)$ for any m, we have:

$$\begin{aligned}&\mathrm{Tr}(\ddot{F}(t)F^*(t)F(t)F^*(t))\\ &\quad= \sum_{m=1}^M \langle \ddot{F}(t)F^*(t)F(t)F^*(t)e_m, e_m\rangle\\ &\quad= \sum_{m=1}^M \langle F^*(t)F(t)f_m(t), \ddot{f}_m(t)\rangle\\ &\quad= \sum_{m=1}^M \langle F^*(t)F(t)f_m(t), -\|g_m\|^2 f_m(t)\rangle\\ &\quad= -\sum_{m=1}^M \|g_m\|^2\|F(t)f_m(t)\|^2.\end{aligned} \tag{7}$$

In particular, combining (7) with Lemma 1 gives:

$$\begin{aligned}\ddot{\varphi}(t) &= -4\sum_{m=1}^M \|g_m\|^2\|F(t)f_m(t)\|^2\\ &\quad + 4\|\dot{F}(t)F^*(t)\|_{\mathrm{HS}}^2\\ &\quad + 2\|\dot{F}^*(t)F(t) + F^*(t)\dot{F}(t)\|_{\mathrm{HS}}^2.\end{aligned} \tag{8}$$

To bound (8), note that by (5),

$$\begin{aligned}\|F(t)\|_{\mathrm{HS}}^2 &= \sum_{m=1}^M \|f_m(t)\|^2 = M,\\ \|\dot{F}(t)\|_{\mathrm{HS}}^2 &= \sum_{m=1}^M \|\dot{f}_m(t)\|^2 = \sum_{m=1}^M \|g_m\|^2,\end{aligned}$$

and thus, taking absolute values of (8), we have:

$$\begin{aligned}
&|\ddot{\varphi}(t)| \\
&\leq 4\sum_{m=1}^{M}\|g_m\|^2\|F(t)f_m(t)\|^2 + 4\|\dot{F}(t)F^*(t)\|_{\mathrm{HS}}^2 \\
&\quad + 2\|\dot{F}^*(t)F(t) + F^*(t)\dot{F}(t)\|_{\mathrm{HS}}^2 \\
&\leq 4\sum_{m=1}^{M}\|g_m\|^2\|F(t)\|_2^2\|f_m(t)\|^2 + 4\|\dot{F}(t)F^*(t)\|_{\mathrm{HS}}^2 \\
&\quad + 2\big(\|\dot{F}^*(t)F(t)\|_{\mathrm{HS}} + \|F^*(t)\dot{F}(t)\|_{\mathrm{HS}}\big)^2 \\
&\leq 4\sum_{m=1}^{M}\|g_m\|^2\|F(t)\|_{\mathrm{HS}}^2 + 12\|\dot{F}(t)\|_{\mathrm{HS}}^2\|F(t)\|_{\mathrm{HS}}^2 \\
&= 16M\sum_{m=1}^{M}\|g_m\|^2. \qquad (9)
\end{aligned}$$

Substituting (7) and (9) into (4) yields (2). □

In light of the Taylor expansion (2), one, in light of Cauchy's inequality, might expect the gradient of FP, namely the $\{g_m\}_{m=1}^M \in \mathbb{H}_N^M$ which maximizes the linear term

$$\mathrm{Re}\sum_{m=1}^{M}\langle F^*Ff_m, g_m\rangle,$$

to be given by $g_m = F^*Ff_m$ for all $m = 1, \ldots, M$. Indeed, one may show that this would be the correct gradient if the frame potential was being regarded as a functional over the entire space $\mathbb{H}_N^M$. However, as we are optimizing over $\mathbb{S}_N^M$, we require that $\{g_m\}_{m=1}^M \in \oplus f_m^{\perp}$, and as such, instead take $\{g_m\}_{m=1}^M$ to be the projection of $\{F^*Ff_m\}_{m=1}^M$ onto $\oplus f_m^{\perp}$. In the next result, we formally verify that such a choice is indeed optimal.

Theorem 3. *For any $\{f_m\}_{m=1}^M \in \mathbb{S}_N^M$, the minimizer of the bound in* (2) *over all $t \in \mathbb{R}$ and $\{g_m\}_{m=1}^M \in \oplus f_m^{\perp}$ is given by $t = -1/(4M)$ and*

$$g_m = F^*Ff_m - \|Ff_m\|^2 f_m, \quad m = 1, \ldots, M.$$

In particular, there exists $\{\tilde{f}_m\}_{m=1}^M \in \mathbb{S}_N^M$ such that:

$$\sum_{m=1}^{M}\|\tilde{f}_m - f_m\|^2 \leq \frac{1}{16M^2}\sum_{m=1}^{M}\big(\|F^*Ff_m\|^2 - \|Ff_m\|^4\big), \quad (10)$$

and such that:

$$\mathrm{FP}(\{\tilde{f}_m\})_{m=1}^M) - \mathrm{FP}(\{f_m\}_{m=1}^M) \leq -\frac{1}{2M}\sum_{m=1}^{M}\big(\|F^*Ff_m\|^2 - \|Ff_m\|^4\big). \quad (11)$$

Proof. We seek to minimize:

$$4t\mathrm{Re}\sum_{m=1}^{M}\langle F^*Ff_m, g_m\rangle + 8Mt^2\sum_{m=1}^{M}\|g_m\|^2 = \frac{2}{M}\sum_{m=1}^{M}\mathrm{Re}\langle F^*Ff_m + 2Mtg_m, 2Mtg_m\rangle \quad (12)$$

over all $\{g_m\}_{m=1}^M \in \mathbb{S}_N^M$ and all $t \in \mathbb{R}$. We note immediately from (12) that the optimal $\{g_m\}_{m=1}^M$ and t are not unique, though we now show that their product is. Indeed, for any fixed m, letting P_m denote the orthogonal projection of $\mathbb{H}_N$ onto the orthogonal complement of f_m, we have:

$$\begin{aligned}
&\mathrm{Re}\langle F^*Ff_m + 2Mtg_m, 2Mtg_m\rangle \\
&= \mathrm{Re}\langle F^*Ff_m + 2Mtg_m, 2MtP_mg_m\rangle \\
&= \mathrm{Re}\langle P_mF^*Ff_m + 2Mtg_m, 2Mtg_m\rangle \\
&= \tfrac{1}{4}\big(\|P_mF^*Ff_m + 4Mtg_m\|^2 - \|P_mF^*Ff_m\|^2\big) \\
&\geq -\tfrac{1}{4}\|P_mF^*Ff_m\|^2,
\end{aligned}$$

with equality if and only if $P_mF^*Ff_m + 4Mtg_m = 0$. Thus, to minimize (12), and consequently to minimize the upper bound in (2), we may take $t = -1/(4M)$ and

$$\begin{aligned}
g_m &= P_mF^*Ff_m \\
&= F^*Ff_m - \langle F^*Ff_m, f_m\rangle f_m \\
&= F^*Ff_m - \|Ff_m\|^2 f_m, \qquad (13)
\end{aligned}$$

as claimed. Moreover, in light of (13), we have:

$$\begin{aligned}
\|g_m\|^2 &= \langle F^*Ff_m, g_m\rangle \\
&= \big\langle F^*Ff_m, F^*Ff_m - \|Ff_m\|^2 f_m\big\rangle \\
&= \|F^*Ff_m\|^2 - \|Ff_m\|^4,
\end{aligned}$$

which, when substituted into (1) and (2) yields (10) and (11), respectively, where $\tilde{f}_m := f_m(-1/4M)$. □

Note that as $\|Ff_m\|^4 = |\langle F^*Ff_m, f_m\rangle|^2 \leq \|F^*Ff_m\|^2$ for all $m = 1, \ldots, M$, Theorem 3 provides a direction and step size in which to travel from a given $\{f_m\}_{m=1}^M \in \mathbb{S}_N^M$ so as to produce a concrete decrease in frame potential. In the next section, we estimate the size of this decrease in terms of how far the current potential is from its minimum, and in so doing, provide an upper bound on the rate at which repeated applications of Theorem 3 will asymptotically produce a unit norm tight frame.

3. Gradient descent of the frame potential

We now consider the gradient descent of the frame potential: by repeatedly applying Theorem 3, we hope to produce a sequence of unit norm frames which are converging to a unit norm tight frame. Here, the main idea is to estimate the right hand side of (11) as a proportion of the difference between the current value of the frame potential and its minimum.

To be clear, in [1], the minimum value of FP over $\mathbb{S}_N^M$ is found to be M^2/N; we now show how the quantity $\mathrm{FP}(\{f_m\}_{m=1}^M) - M^2/N$ is a good metric on the tightness of $\{f_m\}_{m=1}^M$. Indeed, letting $\{\lambda_n\}_{n=1}^N$ be the eigenvalues of the corresponding frame operator F^*F, we have:

$$\sum_{n=1}^{N}\lambda_n = \mathrm{Tr}(F^*F) = \mathrm{Tr}(FF^*) = \sum_{m=1}^{M}\|f_m\|^2 = M. \quad (14)$$

In particular, (14) implies that $\{f_m\}_{m=1}^M \in \mathbb{S}_N^M$ is tight if and only if $\lambda_n = \frac{M}{N}$ for all $n = 1, \ldots, N$. Moreover, as

$$\mathrm{FP}(\{f_m\}_{m=1}^M) = \|FF^*\|_{\mathrm{HS}}^2 = \mathrm{Tr}\big[(F^*F)^2\big] = \sum_{n=1}^N \lambda_n^2,$$

another consequence of (14) is that:

$$\begin{aligned}\mathrm{FP}(\{f_m\}_{m=1}^M) &= \sum_{n=1}^N (\lambda_n - \tfrac{M}{N} + \tfrac{M}{N})^2 \\ &= \sum_{n=1}^N (\lambda_n - \tfrac{M}{N})^2 + 2(0) + \tfrac{M^2}{N},\end{aligned}$$

and thus:

$$\mathrm{FP}(\{f_m\}_{m=1}^M) - \tfrac{M^2}{N} = \sum_{n=1}^N (\lambda_n - \tfrac{M}{N})^2. \qquad (15)$$

That is, the difference between the frame potential and its minimum is the square of the distance of the eigenvalues of F^*F from their optimal values. Using this fact, one may show:

Theorem 4. *For any $\{f_m\}_{m=1}^M \in \mathbb{S}_N^M$,*

$$\begin{aligned}\sum_{m=1}^M \big(\|F^*Ff_m\|^2 - \|Ff_m\|^4\big) \\ \geq \delta^2\big(\mathrm{FP}(\{f_m\}_{m=1}^M) - \tfrac{M^2}{N}\big), \quad (16)\end{aligned}$$

where δ is defined as:

$$\delta := \inf \max_m \min_n |\langle f_m, e_n\rangle|, \qquad (17)$$

where the infimum is taken over all orthonormal bases $\{e_n\}_{n=1}^N$ of $\mathbb{H}_N$.

For sake of space, we omit the complete proof of Theorem 4; the main idea is to let $\{e_n\}_{n=1}^N$ be an orthonormal eigenbasis of F^*F, and note that for any $m = 1, \ldots, M$,

$$\begin{aligned}&\|F^*Ff_m\|^2 - \|Ff_m\|^4 \\ &= \Big\|F^*F\sum_{n=1}^N \langle f_m, e_n\rangle e_n\Big\|^2 \\ &\quad - \Big|\Big\langle F^*F\sum_{n=1}^N \langle f_m, e_n\rangle e_n, f_m\Big\rangle\Big|^2 \\ &= \Big\|\sum_{n=1}^N \lambda_n\langle f_m, e_n\rangle e_n\Big\|^2 - \Big|\sum_{n=1}^N \langle f_m, e_n\rangle\langle \lambda_n e_n, f_m\rangle\Big|^2 \\ &= \sum_{n=1}^N \lambda_n^2|\langle f_m, e_n\rangle|^2 - \Big|\sum_{n=1}^N \lambda_n|\langle f_m, e_n\rangle|^2\Big|^2 \qquad (18) \\ &= \sum_{n=1}^N \Big|\lambda_n - \sum_{p=1}^N \lambda_p|\langle f_m, e_p\rangle|^2\Big|^2|\langle f_m, e_n\rangle|^2, \qquad (19)\end{aligned}$$

where the equality of (18) and (19) arises from the fact that they both represent the variance of the random variable $\{\lambda_n\}_{n=1}^N$ with respect to the probability density function $\{|\langle f_m, e_n\rangle|^2\}_{n=1}^N$.

The significance of Theorem 4 is that it bounds the decrease in frame potential given in Theorem 3 in terms of (15), that is, how far $\{f_m\}_{m=1}^M \in \mathbb{S}_N^M$ is from being tight. Indeed, using Theorem 4, one may show:

Theorem 5. *For any $\{f_m\}_{m=1}^M \in \mathbb{S}_N^M$, there exists $\{\tilde{f}_m\}_{m=1}^M \in \mathbb{S}_N^M$ such that:*

$$\sum_{m=1}^M \|\tilde{f}_m - f_m\|^2 \leq \tfrac{N+1}{16M}\big(\mathrm{FP}(\{f_m\}_{m=1}^M) - \tfrac{M^2}{N}\big), \quad (20)$$

and such that:

$$\begin{aligned}&\mathrm{FP}(\{\tilde{f}_m\})_{m=1}^M) - \tfrac{M^2}{N} \\ &\qquad \leq \big(1 - \tfrac{\delta^2}{2M}\big)\big(\mathrm{FP}(\{f_m\}_{m=1}^M) - \tfrac{M^2}{N}\big), \quad (21)\end{aligned}$$

where δ is given in (17)*.*

By repeatedly applying Theorem 5, one produces a sequence of unit norm frames whose tightness, measured in terms of (15), improves at a geometric rate, provided all δ's remain above some positive lower bound; finding such a bound is a subject of current research.

4. Acknowledgments

Casazza and Fickus were supported by NSF DMS 0704216 and AFOSR F1ATA07337J001, respectively. The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

References:

[1] J.J. Benedetto and M. Fickus. *Finite normalized tight frames*. Adv. Comput. Math., 18:357–385, 2003.

[2] I. Bengtsson and H. Granström. *The frame potential, on average*. Preprint.

[3] P.G. Casazza and M. Fickus. *Minimizing fusion frame potential*. To appear in Acta Appl. Math.

[4] P.G. Casazza, M. Fickus, J. Kovačević, M. Leon and J. Tremain. A physical interpretation of tight frames. In C. Heil, editor, *Harmonic analysis and applications*, pp. 51–76, 2006.

[5] M. Fickus, B.D. Johnson, K. Kornelson, and K. Okoudjou. *Convolutional frames and the frame potential*. Appl. Comput. Harmon. Anal., 19:77–91, 2005.

[6] B.D. Johnson and K. Okoudjou. *Frame potential and finite abelian groups*. Contemp. Math., 464:137–148, 2008.

[7] P. Massey. *Optimal reconstruction systems for erasures and for the q-potential*. Preprint.

[8] P. Massey and M. Ruiz. *Minimization of convex functionals over frame operators*. To appear in Adv. Comput. Math.

[9] P. Massey, M. Ruiz and D. Stojanoff. *The structure of minimizers of the frame potential of fusion frames*. Submitted.

Gabor frames with reduced redundancy

Ole Christensen [(1)], Hong Oh Kim [(2)] and Rae Young Kim [(3)]

(1) Department of Mathematics, Technical University of Denmark, Building 303, 2800 Lyngby, Denmark.
(2) Department of Mathematical Sciences, KAIST, Daejeon, Korea.
(3) Department of Mathematics, Yeungnam University, Gyeongsan-si,Korea.
Ole.Christensen@mat.dtu.dk, kimhong@kaist.edu, rykim@ynu.ac.kr

This work was supported by the Korea Science and Engineering Foundation (KOSEF) Grant funded by the Korea Government(MOST)(R01-2006-000-10424-0) and by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2006-331-C00014).

Abstract:

Considering previous constructions of pairs of dual Gabor frames, we discuss ways to reduce the redundancy. The focus is on B-spline type windows.

1. Introduction

We will consider Gabor systems in $L^2(\mathbb{R})$, i.e., families of functions $\{E_{mb}T_n g\}_{m,n\in\mathbb{Z}}$, where

$$E_{mb}T_n g(x) := e^{2\pi imbx} g(x - na).$$

If there exists a constant $B > 0$ such that

$$\sum_{m,n\in\mathbb{Z}} |\langle f, E_{mb}T_n g\rangle|^2 \leq B\,||f||^2,\ \forall f \in L^2(\mathbb{R}),$$

then $\{E_{mb}T_n g\}_{m,n\in\mathbb{Z}}$ is called a Bessel sequence. If there exist two constants $A, B > 0$ such that

$$A\,||f||^2 \leq \sum_{m,n\in\mathbb{Z}} |\langle f, E_{mb}T_n g\rangle|^2 \leq B\,||f||^2,\ \forall f \in L^2(\mathbb{R}),$$

then $\{E_{mb}T_n g\}_{m,n\in\mathbb{Z}}$ is called a frame. If $\{E_{mb}T_n g\}_{m,n\in\mathbb{Z}}$ is a frame with dual frame $\{E_{mb}T_n h\}_{m,n\in\mathbb{Z}}$, then

$$f = \sum_{m,n\in\mathbb{Z}} \langle f, E_{mb}T_n h\rangle E_{mb}T_n g,\ f \in L^2(\mathbb{R}),$$

where the series expansion converges unconditionally in $L^2(\mathbb{R})$.
Our starting point is the duality condition for Gabor frames, originally due to Ron and Shen [4]. We use the version due to Janssen [3]:

Lemma 1..1 *Two Bessel sequences $\{E_{mb}T_n g\}_{m,n\in\mathbb{Z}}$ and $\{E_{mb}T_n h\}_{m,n\in\mathbb{Z}}$ form dual Gabor frames for $L^2(\mathbb{R})$ if and only if*

$$\sum_{k\in\mathbb{Z}} \overline{g(x - n/b + k)}h(x + k) \quad = \quad b\delta_{n,0} \qquad (1..1)$$

for a.e. $x \in [0, 1]$.

The Bessel condition in Lemma 1..1 is always satisfied for bounded windows with compact support, see [1]. Note that if g and h have compact support, we only need to check a finite number of conditions in (1..1). In this paper we will usually choose b so small that only the condition for $n = 0$ has to be verified.

2. The range $\frac{1}{2N-1} < b < \frac{1}{N}$

We first cite a result from [2]. It yields an explicit construction of dual Gabor frames:

Theorem 2..1 *Let $N \in \mathbb{N}$. Let $g \in L^2(\mathbb{R})$ be a real-valued bounded function with supp $g \subset [0, N]$, for which*

$$\sum_{n\in\mathbb{Z}} g(x - n) = 1. \qquad (2..1)$$

Let $b \in]0, \frac{1}{2N-1}]$. Consider any scalar sequence $\{a_n\}_{n=-N+1}^{N-1}$ for which

$$a_0 = b \text{ and } a_n + a_{-n} = 2b,\ n = 1, 2, \cdots N - 1, \qquad (2..2)$$

and define $h \in L^2(\mathbb{R})$ by

$$h(x) = \sum_{n=-N+1}^{N-1} a_n g(x + n). \qquad (2..3)$$

Then g and h generate dual frames $\{E_{mb}T_n g\}_{m,n\in\mathbb{Z}}$ and $\{E_{mb}T_n h\}_{m,n\in\mathbb{Z}}$ for $L^2(\mathbb{R})$.

The above result can be extended:

Corollary 2..2 *Consider any $b \leq 1/N$. With g and a_n as in Theorem 2..1, the function*

$$h(x) = \left(\sum_{n=-N+1}^{N-1} a_n g(x + n)\right) \chi_{[0,N]}(x) \qquad (2..4)$$

is a dual frame generator of g.

Proof. Consider the condition (1..1) for $n = 0$; only the values of $h(x)$ for $x \in [0, N]$ play a role, so since the condition holds for the function in (2..3), it also holds for the function in (2..4). □

The cut-off in (2..4) yields a non-smooth function. However, for any $b < 1/N$, we might modify h slightly and obtain a smooth dual generator:
In particular, we obtain the following:

Corollary 2..3 *Consider any $b < 1/N$, and take $\epsilon < 1/b - N$. With g as in Theorem 2..1, the function $h(x) = b, x \in [0, N]$ has an extension to a function of desired smoothness, supported on $[-\epsilon, N + \epsilon]$, which is a dual frame generator of g.*

Proof. The choice $a_n = b,\ n = -N+1, \dots, N-1$, leads to

$$\sum_{n=-N+1}^{N-1} a_n g(x+n) = b,\ x \in [0, N].$$

Given $\epsilon < 1/b - N$ and any functions $\phi_1 : [-\epsilon, 0[\to \mathbb{R}$ and $\phi_2 :]N, N+\epsilon] \to \mathbb{R}$, the function

$$h(x) = \begin{cases} \phi_1(x), & x \in [-\epsilon, 0[, \\ \sum_{n=-N+1}^{N-1} a_n g(x+n) = b, & x \in [0, N], \\ \phi_2, & x \in]N, N+\epsilon], \\ 0, & x \notin [-\epsilon, N+\epsilon], \end{cases}$$

will satisfy (1..1); in fact, for $n \neq 0$, the support of the functions $g(\cdot \pm n/b)$ and h are disjoint, and for $n = 0$ we are (for all relevant values of x) back at the function in (2..4). The functions ϕ_1 and ϕ_2 can be chosen such that the function h has the desired smoothness. □

The assumptions in Theorem 2..1 are tailored to B-splines, defined inductively by

$$B_1 := \chi_{[0,1]},\ \ B_{N+1} := B_N * B_1.$$

Direct calculations shows that

$$B_2(x) \ = \ \begin{cases} x & \text{if } x \in [0,1], \\ 2-x & \text{if } x \in [1,2], \\ 0 & \text{otherwise,} \end{cases}$$

and

$$B_3(x) \ = \ \begin{cases} \frac{1}{2}x^2 & \text{if } x \in [0,1], \\ -x^2 + 3x - \frac{3}{2} & \text{if } x \in [1,2], \\ \frac{1}{2}x^2 - 3x + \frac{9}{2} & \text{if } x \in [2,3], \\ 0 & \text{otherwise.} \end{cases}$$

In general, the functions B_N are $(N-2)-$times differentiable piecewise polynomials (explicit expressions are known). Furthermore, supp $B_N = [0, N]$, and the partition of unity condition (2..1) is satisfied.
In case $g = B_N$, the dual generators in Theorem 2..1 are splines, of the same smoothness as B_N itself. By compressing the function $\sum_{n=-N+1}^{N-1} a_n g(x+n)$ from the interval $[-N+1, 0]$ to $[-\epsilon, 0]$ and from $[N, 2N-1]$ to $[N, N+\epsilon]$ we obtain a dual in (2..3) with the same features:

Example 2..4 For the B-spline $B_3(x)$ and $b = 1/5$, Theorem 2..1 yields the symmetric dual

$$h_3(x) = \frac{1}{5} \begin{cases} 1/2\, x^2 + 2\, x + 2, & x \in [-2, -1[, \\ -1/2\, x^2 + 1, & x \in [-1, 0[, \\ 1, & x \in [0, 3[, \\ -1/2\, x^2 + 3\, x - 7/2, & x \in [3, 4[, \\ 1/2\, x^2 - 5x + 25/2, & x \in [4, 5[, \\ 0, & x \notin [0, 5[. \end{cases} \tag{2..5}$$

See Figure 1.
Now, for $b = 1/4$, we can use Corollary 2..3 for $\epsilon < 4 - 3 = 1$. Taking $\epsilon = 1/2$, we compress the function

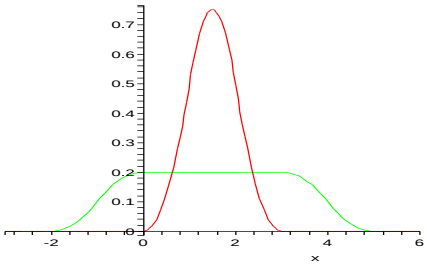


Figure 1: B_3 and the dual generator h_3 in (2..5).

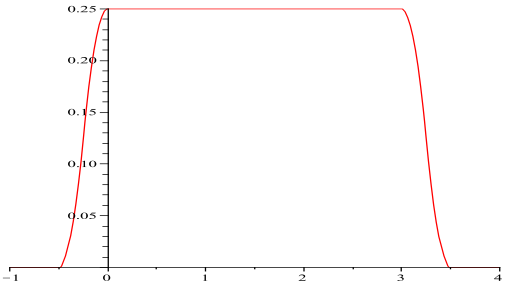


Figure 2: The function h in (3..13)..

h_3 in (2..5) from $[-2, 0]$ to $[-1/2, 0]$ and from $[3, 5]$ to $[3, 31/2]$ and obtain the dual

$$h(x) = \frac{1}{4} \begin{cases} 1/2\,(4x)^2 + 2\,(4x) + 2, & x \in [-1/2, -1/4[, \\ -1/2\,(4x)^2 + 1, & x \in [-1/4, 0[, \\ 1, & x \in [0, 3[, \\ -1/2\,(4(x-3)+3)^2 + 3\,(4(x-3)+3) - 7/2, & \\ & x \in [3, 3+1/4[, \\ 1/2\,(4(x-3)+3)^2 - 5(4(x-3)+3) + 25/2, & \\ & x \in [3+1/4, 3+1/2[, \\ 0, & x \notin [-1/2, 3+1/2[. \end{cases}$$

$$= \frac{1}{4} \begin{cases} 8\,x^2 + 8\,x + 2, & x \in [-1/2, -1/4[, \\ -8\,x^2 + 1, & x \in [-1/4, 0[, \\ 1, & x \in [0, 3[, \\ -8\,x^2 + 48\,x - 71, & x \in [3, 3+1/4[, \\ 8\,x^2 - 56\,x + 98, & x \in [3+1/4, 3+1/2[, \\ 0, & x \notin [-1/2, 3+1/2[. \end{cases}$$

See Figure 2. □

3. B_2 and $1/2 < b < 1$

In the following discussion, we consider dual windows associated with a Gabor frame $\{E_{mb} T_n B_2\}_{m,n \in \mathbb{Z}}$ generated by the B-spline B_2. The arguments can be extended to general functions supported on $[0, 2]$. Take any function h with values specified only on $[0, 2]$ and such that

$$\sum_{k \in \mathbb{Z}} B_2(x+k) h(x+k) = 1,\ x \in [0, 1]. \tag{3..1}$$

In fact, due to the support of B_2, only the values for $h(x)$ for $x \in [0, 2]$ play a role for that condition. We know that

for any $b \leq 1/2$ the function generates – up to a certain scalar multiple – a dual of g.
Now consider any $1/2 < b < 1$; that is, we have $1 < 1/b < 2$.

Lemma 3..1 *Assume that $h(x),\ x \in [0,2]$ is chosen such that* (3..1) *is satisfied. The the following hold:*

(i) If

$$\sum_{k\in\mathbb{Z}} B_2(x-1/b+k)h(x+k) = 0,\ x \in \mathbb{R}, \quad (3..2)$$

and

$$\sum_{k\in\mathbb{Z}} B_2(x+1/b+k)h(x+k) = 0,\ x \in \mathbb{R}, \quad (3..3)$$

then

$$B_2(x-1/b)h(x) + B_2(x-1/b+1)h(x+1) = 0,$$

$$x \in [1/b, 2], \quad (3..4)$$

$$B_2(x+1/b-1)h(x-1) + B_2(x+1/b)h(x) = 0$$

$$x \in [0, 2-1/b]. \quad (3..5)$$

These equations determine $h(x)$ for

$$x \in [-1, 1-1/b] \cup [1+1/b, 3].$$

(ii) If $h(x)$ for $x \in [-1, 1-1/b] \cup [1+1/b, 3]$ is chosen such that (3..4) *and* (3..5) *are satisfied, and*

$$h(x) = 0, x \notin [0,2] \cup [-1, 1-1/b] \cup [1+1/b, 3],$$

then (3..2) *and* (3..3) *hold.*

Proof. We consider (3..2) for $x \in [1,2]$, and split into two cases:
For $x \in [1, 1/b]$, (3..2) yields that

$$\begin{aligned} 0 \ &= \ B_2(x-1/b+1)h(x+1) \\ &\quad + B_2(x-1/b+2)h(x+2); \end{aligned} \quad (3..6)$$

the equation only involve $h(x)$ for

$$x \in [2, 1+1/b] \cup [3, 2+1/b].$$

For $x \in [1/b, 2]$, (3..2) yields that

$$0 = B_2(x-1/b)h(x) + B_2(x-1/b+1)h(x+1);$$

since $h(x)$ is known, this implies that

$$h(x+1) = \frac{-B_2(x-1/b)h(x)}{B_2(x-1/b+1)},\ x \in [1/b, 2],$$

that is,

$$h(x) = \frac{-B_2(x-1/b-1)h(x-1)}{B_2(x-1/b)},\ x \in [1/b+1, 3].$$

Similarly, considering (3..3) for

$$x \in [0,1] = [0, 2-1/b] \cup [2-1/b, 1]$$

leads to (3..5) and

$$B_2(x+1/b-2)h(x-2) + B_2(x+1/b-1)h(x-1)$$

$$= 0,\ x \in [2-1/b, 1]; \quad (3..7)$$

the equation (3..7) only involves $h(x)$ for

$$x \in [-1/b, -1] \cup [1-1/b, 0],$$

and (3..5) implies that

$$h(x-1) = \frac{-B_2(x+1/b)h(x)}{B_2(x+1/b-1)},\ x \in [0, 2-1/b],$$

i.e.,

$$h(x) = \frac{-B_2(x+1/b+1)h(x+1)}{B_2(x+1/b)},\ x \in [-1, 1-1/b].$$

For the proof of (ii), the condition

$$h(x) = 0, x \notin [0,2] \cup [-1, 1-1/b] \cup [1+1/b, 3],$$

implies that (3..6) and (3..7) are satisfied. By construction, (3..2) and (3..3) are satisfied. □

Lemma 3..1 shows that if we want that (3..1), (3..2), and (3..3) hold for some $b \in]1/2, 1]$, then h in general will take values outside $[0,2]$. However, the proof shows that we under certain circumstances can find a solution h having support in $[0,2]$. In that case, the support will actually be a subset of $[0,2]$:

Corollary 3..2 *Let $b \in]1/2, 1]$. Assume that supp $h \subseteq [0,2]$ and that* (3..1) *and* (3..2) *holds. Then*

$$h(x) = 0,\ x \in [0, 2-1/b] \cup [1/b, 2]. \quad (3..8)$$

Proof. According to the proof of Lemma 3..1, we obtain that $h(x) = 0$ on $[1/b+1, 3]$ by requiring that $h(x) = 0$ for $x \in [1/b, 2]$; and we obtain that $h(x) = 0$ on $[-1, 1-1/b]$ by requiring that $h(x) = 0$ for $x \in [0, 2-1/b]$. □

If supp $h \subseteq [0,2]$, the condition (3..8) implies that h at most can be nonzero on the interval $[2-1/b, 1/b]$ having length $2/b - 2$. In order for (3..1) to hold, this interval must have length at least 1; thus, we need to consider b such that $2/b - 2 \geq 1$, i.e., $b \leq 2/3$. Note that if $b \leq 2/3$, then $2/b \geq 3$: that is, because B_2 and h are supported on $[0,2]$, Janssen's duality conditions in (1..1) are automatically satisfied for $n = \pm 2, \pm 3, \ldots$.

Corollary 3..3 *Consider $b \in]1/2, 2/3]$. Then there exists a function h with supp $h \subseteq [0,2]$ such that* (3..1) *and* (3..2) *hold; and $bh(x)$ is a dual generator of B_2 for these values of b.*

Proof. For $x \in [0, 2-1/b] \cup [1/b, 2]$, let $h(x) = 0$. For $x \in [0,1]$, the equation (3..1) means that

$$xh(x) + (1-x)h(x+1) = 1.$$

This implies that

$$\begin{aligned} xh(x) &= 1,\ x \in [1/b-1, 1], \\ (1-x)h(x+1) &= 1,\ x \in [0, 2-1/b]; \end{aligned}$$

that is,

$$h(x) = \frac{1}{x},\ x \in [1/b-1, 1], \tag{3..9}$$

and

$$h(x) = \frac{1}{2-x},\ x \in [1, 3-1/b]. \tag{3..10}$$

Finally, for $x \in [2-1/b, 1/b-1]$ and $x \in [3-1/b, 1/b]$, choose $h(x)$ such that

$$xh(x) + (1-x)h(x+1) = 1.$$

By construction, $bh(x)$ is a dual generator. □

For $b = 3/5$ we will now explicitly construct a continuous dual generator h of B_2 with support in $[0,2]$. Putting Corollary 3..2, (3..9), and (3..10) together, we can state a result about how a dual window supported on $[0,2]$ must look like on parts of $[0,2]$:

Lemma 3..4 *For $b = 3/5$, every dual generator of B_2 with support in $[0,2]$ has the form*

$$h(x) = \begin{cases} 0 & \text{if } x \le 1/3; \\ \frac{1}{x} & \text{if } x \in [2/3, 1]; \\ \frac{1}{2-x} & \text{if } x \in [1, 4/3]; \\ 0 & \text{if } x \ge 5/3. \end{cases}$$

That is, we only have freedom on the definition of h on $]1/3, 2/3[\cup]4/3, 5/3[$.

Note that on $[2/3, 4/3]$, the function h is symmetric around $x = 1$. We will now show that it is possible to define h on $]1/3, 2/3[\cup]4/3, 5/3[$ in such a way that h becomes symmetric around $x = 1$.
First, we note that this form of symmetry means that

$$h(1-x) = h(1+x),\ x \in]1/3, 2/3[. \tag{3..11}$$

Put together with the duality condition, we thus require that

$$xh(x) = 1 - (1-x)h(1-x),\ x \in]1/3, 2/3[. \tag{3..12}$$

The condition (3..12) shows that must define $h(1/2) = 1$. Now, taking any continuous function h defined on $[1/3, 1/2]$ with the properties that $h(1/3) = 0$ and $h(1/2) = 1$, the condition (3..12) shows how to define $h(x)$ on $]1/2, 2/3[$; and, finally, the condition (3..11) shows how to define h on $]4/3, 5/3[$ such that the resulting function is a symmetric dual generator.

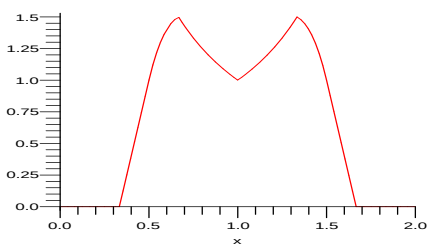


Figure 3: The function h in (3..13)..

Put

$$h(x) = 6x - 2,\ x \in [1/3, 1/2].$$

Then, for $x \in [1/2, 2/3]$,

$$\begin{aligned} h(x) &= \frac{1-(1-x)h(1-x)}{x} \\ &= \frac{-6x^2+10x-3}{x}. \end{aligned}$$

The condition $h(1+x) = h(1-x),\ x \in]1/3, 2/3[$ can also be expressed as $h(x) = h(2-x),\ x \in]4/3, 5/3[$. Thus, for $x \in [4/3, 3/2]$ we arrive at

$$h(x) = h(2-x) = \frac{-6x^2+14x-7}{2-x},\ x \in [4/3, 3/2];$$

while, for $x \in [3/2, 5/3]$,

$$h(x) = h(2-x) = 6(2-x) - 2 = 10 - 6x.$$

We have arrived at the following conclusion:

Lemma 3..5 *For $b = 3/5$, the function*

$$h(x) = \begin{cases} 0 & \text{if } x \le 1/3; \\ 6x-2 & \text{if } x \in [1/3, 1/2]; \\ \frac{-6x^2+10x-3}{x} & \text{if } x \in [1/2, 2/3]; \\ \frac{1}{x} & \text{if } x \in [2/3, 1]; \\ \frac{1}{2-x} & \text{if } x \in [1, 4/3]; \\ \frac{-6x^2+14x-7}{2-x} & \text{if } x \in [4/3, 3/2]; \\ 10-6x & \text{if } x \in [3/2, 5/3]; \\ 0 & \text{if } x \ge 5/3 \end{cases} \tag{3..13}$$

is a continuous symmetric dual generator of B_2.

References:

[1] Christensen, O.: *Frames and bases. An introductory course.* Birkhäuser 2007.

[2] Christensen, O. and Kim, R. Y.: *On dual Gabor frame pairs generated by polynomials.* J. Fourier Anal. Appl., accepted for publication.

[3] Janssen, A.J.E.M.: *The duality condition for Weyl-Heisenberg frames.* In "Gabor analysis: theory and applications" (eds. H.G. Feichtinger and T. Strohmer). Birkhäuser, Boston, 1998.

[4] Ron, A. and Shen, Z.: *Frames and stable bases for shift-invariant subspaces of $L^2(\mathbb{R}^d)$.* Canad. J. Math. **47** no. 5 (1995), 1051–1094.

Linear independence and coherence of Gabor systems in finite dimensional spaces

Götz E. Pfander [(1)],

(1) Jacobs University, 28759 Bremen, Germany.
g.pfander@jacobs-university.de

Abstract:

This paper reviews recent results on the geometry of Gabor systems in finite dimensions. For example, we discuss the coherence of Gabor systems, the linear independence of subsets of Gabor systems, and the condition number of matrices formed by a small number of vectors from a Gabor system. We state a result on the recovery of signals that have a sparse representation in certain Gabor systems. The results listed here are obtained by the author in collaborations with Jim Lawrence, Felix Krahmer, Peter Rashkov, Jared Tanner, Holger Rauhut, and David Walnut linear independence

1. Introduction and Notation

The theory of Gabor systems in the Hilbert space of square integrable functions on the real line has received significant attention during the last ten to twenty years (see, for example, [4, 6, 8, 7] and references within). Much of the research concentrates on showing that certain Gabor systems are frames or Riesz bases for their closed linear span. The seemingly simpler concept of linear independence of vectors in a Gabor system was addressed in [10]. There, it was conjectured that any finite set of time–frequency shifted copies of a single square integrable function is linear independent. This conjecture still remains to be resolved.

In the last years, in part due to the emergence of the theory of compressed sensing and sparse signal recovery, the structure of Gabor systems in finite dimensional spaces has received increased attention. Such finite Gabor systems on finite Abelian groups are described below.

We let G denote a finite Abelian group. Its dual group $\widehat{G}$ consists of the group homomorphisms $\xi : G \mapsto S^1$. We have $\widehat{G} \subseteq \mathbb{C}^G = \{f : G \longrightarrow \mathbb{C}\}$, the latter being the space of complex valued functions on G. The support size of $f \in \mathbb{C}^G$ is $\|f\|_0 := |\{x : f(x) \neq 0\}|$. The Fourier transform of $f \in \mathbb{C}^G$ is normalized to be $\widehat{f}(\xi) = \sum_{x\in G} f(x)\,\overline{\xi(x)}$, $\xi \in \widehat{G}$.

Translation operators T_x, $x \in G$, and modulation operators M_ξ, $\xi \in \widehat{G}$, on $\mathbb{C}^G$ are unitary operators given by $(T_x f)(t) = f(t-x)$ and $(M_\xi f)(t) = f(t)\cdot\xi(t)$. Time-frequency shift operators $\pi(\lambda)$, $\lambda = (x,\xi) \in G \times \widehat{G}$, are the unitary operator on $\mathbb{C}^G$ represented by $\pi(\lambda)f = T_x \circ M_\xi f$, $\lambda = (x,\xi) \in G \times \widehat{G}$.

The system $\{\pi(\lambda)g : \lambda \in G \times \widehat{G}\} \subseteq \mathbb{C}^G$ is called (full) Gabor system with window $g \in \mathbb{C}^G$, it consists of $|G|^2$ vectors in a $|G|$ dimensional space.

The short-time Fourier transform with respect to g is given by

$$V_g f(\lambda) \quad = \quad \langle f, \pi(\lambda)g\rangle = \sum_{y\in G} f(y)\overline{g(y-x)\xi(y)},$$
$$f \in \mathbb{C}^G, \lambda = (x,\xi) \in G \times \widehat{G}.$$

We shall not make a distinction between the linear mapping $V_g : \mathbb{C}^G \longrightarrow \mathbb{C}^{G\times\widehat{G}}$ and its matrix representation with respect to the Euclidean basis.

Full Gabor systems in finite dimensions share an important and very useful property: for any $g \neq 0$, the collection $\{\pi(\lambda)g\}_{\lambda\in G\times\widehat{G}}$ forms a uniform tight finite frame for $\mathbb{C}^G$ with frame bound $n^2\,\|g\|^2$, that is,

$$\sum_{\lambda\in G\times\widehat{G}} |\langle f, \pi(\lambda)g\rangle|^2 = n^2\,\|g\|^2\|f\|^2.$$

This is a simple consequence of the representation theory of the Weyl–Heisenberg group [9, 12].

In this paper we are concerned with properties of subsets of full Gabor systems. In Section 2, we consider the linear independence of subsets of $|G|$ elements of $\{\pi(\lambda)g\}_{\lambda\in G\times\widehat{G}}$. Recall that a finite set of vectors in $\mathbb{C}^G$ is in general linear position if any subset of at most $|G|$ of these vectors are linearly independent. While being a classical concept in mathematics, it is also relevant for communications, namely, for information transmission through a so-called erasure channel [2]. In fact, a frame $\mathcal{F} = \{x_k\}_{k=1}^m$ in $\mathbb{C}^n$ is called maximally robust to erasures if the removal of any $l \leq m-n$ vectors from $\mathcal{F}$ leaves a frame.

Moreover, we consider the coherence of Gabor systems in Section 3. We state probabilistic estimates of the coherence of a full Gabor system with respect to a randomly generated window. In Section 4, we consider the condition number of matrices formed by a small subset of a Gabor system.

The results presented below were obtained over the last few years in collaboration with Jim Lawrence and David Walnut [12], Felix Krahmer and Peter Rashkov [11], and Holger Rauhut and Jared Tanner [14, 13].

2. Gabor systems in general linear position

The following simple observations illustrate the usefulness of Gabor systems which are in general linear position.

Proposition 1 *[11, 12] For $g \in \mathbb{C}^G \setminus \{0\}$, the following are equivalent:*

1. *$\{\pi(\lambda)g\}_{\lambda\in G\times\widehat{G}}$ are in general linear position.*
2. *For all $f \in \mathbb{C}^G\backslash\{0\}$ we have $\|V_g f\| \geq |G|^2-|G|+1$.*
3. *For all $f \in \mathbb{C}^G$, $V_g f$ is completely determined by its values on any set Λ with $|\Lambda| = n$.*
4. *$\{\pi(\lambda)g\}_{\lambda\in G\times\widehat{G}}$ is maximally robust to erasures.*
5. *The $|G| \times |G|^2$ matrix V_g has the property that every minor of order n is nonzero.*

Corollary 2 *[12] If $\{\pi(\lambda)g\}_{\lambda\in G\times\widehat{G}}$ are in general linear position, then $\|g\|_0 = |G|$ and $\|\widehat{g}\|_0 = |G|$.*

Unfortunately, not each finite Abelian groups G permits the existence of a vector $g \in \mathbb{C}^G$ satisfying one and therefore all conditions listed in Proposition 1. For example, for the group $G = \mathbb{Z}_2 \times \mathbb{Z}_2$, no such g exists [11]. The situation is different for $G = \mathbb{Z}_p$. Recall that E is of full measure if the Lebesgue measure of $\mathbb{C}^G \setminus E$ is 0.

Theorem 3 *[12] If $|G|$ is prime, that is, $G = \mathbb{Z}_p$, p prime, then there is a dense open set E of full measure in $\mathbb{C}^G$ such that for every $g \in E$, the elements of the full Gabor system $\{\pi(\lambda)g\}_{\lambda\in G\times\widehat{G}}$ are in general linear position. That is, for almost all g we have $\|V_g f\| \geq |G|^2-|G|+1$ for all $f \neq 0$.*

Rudimentary numerical experiments encourage us to ask the following question.

Question 4 *[12] For G cyclic, that is, $G = \mathbb{Z}_n$, $n \in \mathbb{N}$, exists $g \in \mathbb{C}^G$ so that the conclusions of Proposition 1, and, therefore, $\|V_g f\| \geq |G|^2 - |G| + 1$, $f \in \mathbb{C}^G$, hold*

In fact, for $|G|$ prime, Theorem 3 can be strengthened.

Theorem 5 *[11] Let $G = \mathbb{Z}_p$, p prime. For almost every $g \in \mathbb{C}^G$, we have*

$$\|V_g f\|_0 \geq |G|^2 - \|f\|_0 + 1 \qquad (1)$$

for all $f \in \mathbb{C}^G \setminus \{0\}$. Moreover, for $1 \leq k \leq |G|$ and $1 \leq l \leq |G|^2$ with $k + l \geq |G|^2 + 1$ there exists f with $\|f\|_0 = k$ and $\|V_g f\|_0 = l$.

Proposition 6 *[11] If $|G|$ is not prime, then V_g has zero minors for all $g \in \mathbb{C}^G$. Hence, there is no $g \in \mathbb{C}^G$ such that* (1) *holds for all $f \in \mathbb{C}^G$.*

Numerical experiments for Abelian groups of order less than or equal to 8, as well as our result for all cyclic groups of prime order, indicate that the following question might have an affirmative answer.

Question 7 *[11] For every cyclic group G and almost every $g \in \mathbb{C}^G$, does*

$$\begin{aligned}&\big\{(\|f\|_0, \|V_g f\|_0),\ f \in \mathbb{C}^G\backslash\{0\}\big\}\\ &= \big\{(\,\|f\|_0\,,\ \|\widehat{f}\|_0+|G|^2-|G|\,),\ f \in \mathbb{C}^G\backslash\{0\}\big\}\end{aligned}$$

hold?

The following result improves on Theorem 5. It allows for the construction of Gabor based equal norm tight frames of p^2 elements in $\mathbb{C}^n$, $n \leq p$. To our knowledge, the only previously known equal norm tight frames that are maximally robust to erasures are so-called harmonic frames (see Conclusions in [2]).

Proposition 8 *[11] There exists a unimodular $g \in \mathbb{C}^{\mathbb{Z}_p}$, p prime, that is, a g with $|g(x)| = 1$ for all $x \in G$ satisfying the conclusions of Theorem 5.*

To construct an equal norm tight frame, we choose a $g \in (S^1)^p$ satisfying the conclusions of Proposition 8. We remove $p - n$ components of the equal norm tight frame $\{\pi(\lambda)g\}_{\lambda\in G\times\widehat{G}}$. The resulting frame remains an equal norm tight frame which is maximally robust to erasure. Note that this frame is not a Gabor frame proper. Reducing the number of vectors in the frame to $m \leq p^2$ vectors leaves an equal norm frame which is maximally robust to erasure but which might not be tight. With the restriction to frames with p^2 elements, p prime, we have shown the existence of Gabor frames which share the usefulness of harmonic frames when it comes to transmission of information through erasure channels.
Background and more details on frames and erasures can be found in [2, 15] and the references cited therein.
Note that Theorem 5 has as direct consequence

Theorem 9 *[11] Let $g \in \mathbb{C}^{\mathbb{Z}_p}$, p prime, satisfy the conclusion of Theorem 5. Then any $f \in \mathbb{C}^{\mathbb{Z}_p}$ with $\|f\|_0 \leq \frac{1}{2}|\Lambda|$, $\Lambda \subset \mathbb{Z}_p\times\widehat{\mathbb{Z}_p}$, is uniquely determined by Λ and $r_\Lambda V_g f$.*

Here, only the support size of f is known. No additional information on the support of f is required to determine f.
In terms of sparse representations, we consider the question whether any vector $f = \sum_{\lambda\in\Lambda} c_\lambda \pi(\lambda)g$ can be determined by a few entries of f in case that $|\Lambda|$ is small.

Theorem 10 *[11] Let $g \in \mathbb{C}^{\mathbb{Z}_p}$, p prime, satisfy the conclusion of Theorem 5. Then any $f \in \mathbb{C}^{\mathbb{Z}_p}$ with $f = \sum_{\lambda\in\Lambda} c_\lambda\pi(\lambda)g$, $\Lambda \subset \mathbb{Z}_p\times\widehat{\mathbb{Z}_p}$ is uniquely determined by B and $r_B f$ whenever $|B| \geq 2|\Lambda|$.*

Note that similar to before, the efficient recovery of f from $2|\Lambda|$ samples of f in Theorem 10 does not require knowledge of Λ.
The question asking how to recover f from a small number of entries of f efficiently will be briefly addressed with Theorem 14

3. Coherence of Gabor systems

In the following we restrict our attention to cyclic groups $G = \mathbb{Z}_n$, $n \in \mathbb{N}$. We consider the so-called Alltop window h^A [15] with entries

$$h^A(x) = \frac{1}{\sqrt{n}} e^{2\pi i x^3/n}, \quad x = 0, \dots, n-1, \qquad (2)$$

and the randomly generated window h^R with entries

$$h^R(x) = \frac{1}{\sqrt{n}} \epsilon_x, \quad x = 0, \dots, n-1, \qquad (3)$$

where the ϵ_x are independent and uniformly distributed on the torus $\{z \in \mathbb{C}, |z| = 1\}$.
For $\|h\|_2 = 1$, the coherence of a full Gabor systems is

$$\mu = \max_{(\ell,p)\neq(\ell',p')} |\langle M_\ell T_p h, M_{\ell'} T_{p'} h\rangle|. \qquad (4)$$

In [16] it is shown that the coherence of $\{\pi(\lambda)h^A : \lambda \in \mathbb{Z}_n \times \widehat{\mathbb{Z}}_n\} \subseteq \mathbb{C}^n$ given in (2) satisfies

$$\mu = \frac{1}{\sqrt{n}} \qquad (5)$$

for n prime. This is close to optimal since as the lower bound for the coherence of frames with n^2 elements in $\mathbb{C}^n$ is $\mu \geq \frac{1}{\sqrt{n+1}}$ [16].
Unfortunately, the coherence (4) of h^A applies only for n prime. For arbitrary n we now consider the random window h^R.

Theorem 11 *[14] Let $n \in \mathbb{N}$ and choose a random window h^R with entries*

$$h^R(x) = \frac{1}{\sqrt{n}} \epsilon_x, \quad x = 0, \dots, n-1,$$

where the ϵ_x are independent and uniformly distributed on the torus $\{z \in \mathbb{C}, |z| = 1\}$. Let μ be the coherence of the associated Gabor dictionary (4), then for $\alpha > 0$ and n even,

$$\mathbb{P}\big(\mu \geq \frac{\alpha}{\sqrt{n}}\big) \leq 4n(n-1)e^{-\alpha^2/4},$$

while for n odd,

$$\mathbb{P}\big(\mu \geq \frac{\alpha}{\sqrt{n}}\big) \leq 2n(n-1)\left(e^{-\frac{n-1}{n}\alpha^2/4} + e^{-\frac{n+1}{n}\alpha^2/4}\right). \qquad (6)$$

Up to the constant factor α, the coherence in Theorem 11 comes close to the lower bound $\mu \geq \frac{1}{\sqrt{n+1}}$ with high probability. (The probability depends on α).

4. Conditioning of submatrices of V_g

For applications such as sparse signal recovery, not only linear independence of subsets of Gabor systems is required. It is rather needed, that small subsets of Gabor systems form well-conditioned matrices.
Throughout this section, we let $\Psi = V_g \in \mathbb{C}^{n\times n^2}$ with $g = h^R$ being the randomly generated unimodular window described in (3). For $\Lambda \subseteq G\times\widehat{G}$ we denote by Ψ_Λ the matrix consisting only of those columns indexed by $\lambda \in \Lambda$.

Theorem 12 *[13] Let $\varepsilon, \delta \in (0,1)$ and $|\Lambda| = S$. Suppose that*

$$S \leq \frac{\delta^2 n}{4e(\log(S/\varepsilon) + c)} \qquad (7)$$

with $c = \log(e^2/(4(e-1))) \approx 0.0724$. Then $\|I_\Lambda - \Psi_\Lambda^*\Psi_\Lambda\| \leq \delta$ with probability at least $1-\varepsilon$; in other words the minimal and maximal eigenvalues of $\Psi_\Lambda^*\Psi_\Lambda$ satisfy $1-\delta \leq \lambda_{\min} \leq \lambda_{\max} \leq 1+\delta$ with probability at least $1-\varepsilon$.

Remark 13 *[13] Assuming equality in condition (7) and solving for ε we deduce*

$$\begin{aligned}\mathbb{P}\big(\|I_\Lambda - \Psi_\Lambda^*\Psi_\Lambda\| > \delta\big) &\leq \frac{e^2}{4(e-1)} S \exp\left(-\frac{\delta^2 n}{4eS}\right) \\ &= CS\exp\left(-\frac{\delta^2 n}{4eS}\right)\end{aligned}$$

with $C \approx 1.075$.

Theorem 12 allows us to guarantee the successful use of efficient algorithms to determine $f = \sum_{\lambda\in\Lambda} c_\lambda \pi(\lambda) g$ from a few entries of f in case that $|\Lambda|$ is small. Here, we will concentrate on algorithms based on Basis Pursuit. Basis Pursuit seeks the solution of the convex problem

$$\min_x \|x\|_1 \quad \text{subject to } \Psi_g x = y, \qquad (8)$$

where $\|x\|_1 = \sum_{\lambda\in\mathbb{Z}_n^2} |x_\lambda|$ is the ℓ_1-norm of x. Efficient convex optimization techniques for Basis Pursuit can be found in [1, 3, 5].

Theorem 14 *[13] Assume x is an arbitrary S-sparse coefficient vector. Choose the random unimodular Gabor window $g = h^R$ defined in (3), that is, with random entries independently and uniformly distributed on the torus $\{z \in \mathbb{C}, |z| = 1\}$. Assume that*

$$S \leq C\frac{n}{\log(n/\varepsilon)} \qquad (9)$$

for some constant C. Then with probability at least $1-\varepsilon$ Basis Pursuit (8) recovers x from $y = \Psi x = \Psi_g x$.

References:

[1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge Univ. Press, 2004.

[2] Peter G. Casazza and Jelena Kovačević. Equal-norm tight frames with erasures. *Adv. Comput. Math.*, 18(2-4):387–430, 2003. Frames.

[3] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by Basis Pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1999.

[4] O. Christensen. *An introduction to frames and Riesz bases*. Applied and Numerical Harmonic Analysis. Birkhäuser Boston Inc., Boston, MA, 2003.

[5] D.L. Donoho and Y. Tsaig. Fast solution of l1-norm minimization problems when the solution may be sparse. *Preprint*, 2006.

[6] H.G. Feichtinger and T. Strohmer, editors. *Gabor Analysis and Algorithms: Theory and Applications*. Birkhäuser, Boston, MA, 1998.

[7] H.G. Feichtinger and T. Strohmer, editors. *Advances in Gabor Analysis*. Applied and Numerical Harmonic Analysis. Birkhäuser Boston Inc., Boston, MA, 2003.

[8] K. Gröchenig. *Foundations of Time-Frequency Analysis*. Applied and Numerical Harmonic Analysis. Birkhäuser, Boston, MA, 2001.

[9] A. Grossmann, J. Morlet, and T. Paul. Transforms associated to square integrable group representations. I. General results. *J. Math. Phys.*, 26(10):2473–2479, 1985.

[10] C. Heil, J. Ramanathan, and P. Topiwala. Linear independence of time–frequency translates. *Proc. Amer. Math. Soc.*, 124(9), September 1996.

[11] F. Krahmer, G.E. Pfander, and P. Rashkov. Uncertainty principles for time–frequency representations on finite abelian groups. *Appl. Comp. Harm. Anal.*, 2008. doi:10.1016/j.acha.2007.09.008.

[12] J. Lawrence, G.E. Pfander, and D. Walnut. Linear independence of Gabor systems in finite dimensional vector spaces. *J. Fourier Anal. Appl.*, 11(6):715–726, 2005.

[13] G.E. Pfander and H. Rauhut. Sparsity in time–frequency representations. 2008. Preprint.

[14] G.E. Pfander, H. Rauhut, and J. Tanner. Identification of matrices having a sparse representation. *IEEE Trans. Signal Proc.*, 2008. to appear.

[15] T. Strohmer and R.W. Heath, Jr. Grassmannian frames with applications to coding and communication. *Appl. Comput. Harmon. Anal.*, 14(3):257–275, 2003.

[16] Thomas Strohmer and Robert W.jun. Heath. Grassmannian frames with applications to coding and communication. *Appl. Comput. Harmon. Anal.*, 14(3):257–275, 2003.

Error Correction for Erasures of Quantized Frame Coefficients

Bernhard G. Bodmann[(1)], Peter G. Casazza[(2)], Gitta Kutyniok[(3)] and Steven Senger[(2)]
(1) Department of Mathematics, University of Houston, Houston, TX 77204, USA.
(2) Department of Mathematics, University of Missouri, Columbia, MO 65211, USA.
(3) Institute of Mathematics, University of Osnabrück, 49069 Osnabrück, Germany.
bgb@math.uh.edu, pete@math.missouri.edu, kutyniok@math.uni-osnabrueck.de,
senger@math.missouri.edu

Abstract:

In this paper we investigate an algorithm for the suppression of errors caused by quantization of frame coefficients and by erasures in their subsequent transmission. The erasures are assumed to happen independently, modeled by a Bernoulli experiment. The algorithm for error correction in this study embeds check bits in the quantization of frame coefficients, causing a possible, but controlled quantizer overload. If a single-bit quantizer is used in conjunction with codes which satisfy the Gilbert Varshamov bound, then the contributions from erasures and quantization to the reconstruction error is shown to have bounds with the same asymptotics in the limit of large numbers of frame vectors.

1. Introduction

The versatility of redundant systems, in particular frames, has been demonstrated by their resilience to erasures and by their usefulness to suppress quantization errors. In the context of finite frames, the statistical error estimates by Goyal, Kovačević, Vetterli and Kelner [7, 6] were to the authors' knowledge the first instance of a combined analysis of erasures and quantization.

In recent years, the robustness of finite frames against erasures has been more extensively studied, for instance, in [5, 14, 9, 2, 10]. These studies typically provide estimates for the (average and worst case) blind reconstruction error, meaning all erased (unknown) coefficients are set to zero and the reconstruction relies on a fixed synthesis operator. It is well-known that if the frame vectors related to the non-erased coefficients still form a spanning set, then the frame operator of those can be inverted, leading to perfect reconstruction. However, the latency caused by the wait until all coefficients have been transmitted and the computational cost of inverting the frame operator make perfect reconstruction less practicable.

On the other hand, Benedetto, Powell and Yilmaz [1] investigated an easily implementable, active error correction for the compensation of quantization errors with so-called sigma-delta algorithms, which provide highly accurate reconstruction.

Recently, Boufounos, Oppenheim and Goyal [4] introduced an erasure correction scheme with strong similarities to quantization-noise shaping, offering the possibility of a combined treatment of both types of errors. The idea of pre-compensation and error-forward projection deserves to be explored further, but the algorithm by Boufounos, Oppenheim and Goyal is computationally still more costly than a simple application of sigma-delta quantization.

The need for results on low-complexity quantization-and-erasure correcting algorithms motivated the present study, which investigates a rather simple strategy for error compensation, a modified sigma-delta algorithm with embedded check bits. The error correction algorithm we present allows precise bounds on quantization errors and also on the effect of erasures from unreliable transmissions of frame coefficients.

2. PCM quantization and blind reconstruction

We first revisit erasure-averaged error bounds for PCM quantization of frame coefficients and blind reconstruction after transmission.

Definition. Let $\mathcal{H}$ be a d-dimensional Hilbert space. A frame $\mathcal{F} = \{f_1, f_2, \dots f_N\}$ for $\mathcal{H}$ is a spanning set. If all vectors in the frame have the same norm, we call $\mathcal{F}$ equal-norm. If $x = \frac{1}{A}\sum_{j=1}^{N}\langle x, f_j\rangle f_j$ for all $x \in \mathcal{H}$, then we say that $\mathcal{F}$ is A-tight.

Quantizing frame coefficients simply means mapping them to a finite set of values.

Definition. A function Q on $\mathbb{R}$ is called *a quantizer with accuracy $\epsilon > 0$ on the interval $[-L, +L]$* if it has a finite range $\mathbb{A}$ and for any $x \in [-L, +L]$, $Q(x)$ satisfies $|x - Q(x)| \leq \epsilon$. The range $\mathbb{A}$ of the quantizer Q is also called the *alphabet*. If this alphabet consists of all integer multiples of a fixed step-size δ contained in the interval $[-L - \delta/2, +L + \delta/2]$ and the quantizer assigns to $x \in [-L, +L]$ the unique value $m\delta, m \in \mathbb{Z}$, satisfying $(m - \frac{1}{2})\delta < x \leq (m + \frac{1}{2})\delta$ then we call Q the *uniform mid-tread quantizer with step-size δ* [3]. Alternatively, if the alphabet is $\mathbb{A} = (\mathbb{Z} + \frac{1}{2})\delta \cap [-L - \delta/2, +L + \delta/2]$ and if Q assigns to $x \in [-L, +L]$ the value $(m+\frac{1}{2})\delta$ such that $m\delta < x \leq (m + 1)\delta$, then we speak of the so-called *uniform mid-riser quantizer with step-size δ*. In the latter part of this study, we focus on the single-bit mid-riser quantizer which rounds the input to $\mathbb{A} = \{-\delta/2, +\delta/2\}$. We want to apply this quantizer to frame coefficients.

Definition. Given a quantizer Q, the *PCM quantization* of a vector x in a real Hilbert space $\mathcal{H}$ of dimension

$\dim(\mathcal{H}) = d$, equipped with an A-tight frame $\mathcal{F} = \{f_j\}_{j=1}^N$, is defined by

$$Q_{\mathcal{F}}(x) = \frac{1}{A}\sum_{j=1}^{N} Q(\langle x, f_j\rangle) f_j\,.$$

Remark. We recall that the PCM quantization error resulting from a uniform quantizer Q with accuracy $\epsilon > 0$ on $[-L, +L]$, and a N/d-tight equal-norm frame $\mathcal{F}$ applied to any input vector $x \in \mathcal{H}$ satisfying $\|x\| \leq L$ is in norm bounded by

$$\begin{aligned}\|Q_{\mathcal{F}}(x) - x\| &\leq \max_{\|v\|=1} \max_{u_j \in \{\pm 1\}} \frac{d}{N} |\sum_{j=1}^{N} u_j \langle f_j, v\rangle| \\ &\leq \frac{d}{N}(\sqrt{N}\epsilon)(\sum_{j=1}^{N} |\langle f_j, v\rangle|^2)^{1/2} = \sqrt{d}\epsilon\,.\end{aligned}$$

This is in contrast to erasures, where the bound on the reconstruction error depends on the norm of the input vector.

Definition. Given a probability measure $\mathbb{P}$ on the set of erasures, and the analysis operator V belonging to an A-tight frame, we define the erasure-averaged reconstruction error to be

$$e(V, \mathbb{P}) = \mathbb{E}[\|\frac{1}{A} V^* E(\omega) V - I\|]\,.$$

Hereby, $\mathbb{E}[\cdot]$ is the expectation with respect to the probability measure $\mathbb{P}$ on $\Omega = \{0, 1\}^N$, and $E : \Omega \to \mathbb{R}^{N\times N}$ is a random diagonal matrix with entries $E_{j,j} = \omega_j$.

Theorem. Let $\mathcal{H}$ be a real Hilbert space of dimension d, equipped with an A-tight equal-norm frame $\mathcal{F}$. If all the frame coefficients are erased with a probability $0 \leq p \leq 1$, independently of each other, then the erasure-averaged reconstruction error is bounded by

$$p \leq \mathbb{E}[\|\frac{1}{A} V^* E(\omega) V - I\|] \leq pd\,.$$

Proof. The lower bound uses Jensen's inequality and the convexity of the norm on the real vector space of Hermitian operators [12]. The upper bound relies on the identity for the operator norms $\|V^*(I - E)V\| = \|(I - E)VV^*(I - E)\|$ and on the bound for entries in the Grammian, $|(VV^*)_{j,k}| \leq \|f_j\|\|f_k\| = 1$, derived from the Cauchy-Schwarz inequality, which implies $\mathbb{E}[\|(I - E(\omega))VV^*(I - E(\omega))\|] \leq Np$. □

Thus, for a vector x for which $p\|x\|$ is bigger than $(\delta/2)\sqrt{d}$, the bound on the worst case error due to erasures dominates that of PCM quantization.

A similar phenomenon happens when the quantization is obtained with first and higher-order sigma delta quantization. For sufficiently large N, the bound for the worst-case quantization error, see e.g. [3], is smaller than the worst-case erasure error. This motivates investigating active error correction for erasures.

3. Sigma-delta quantization with embedded check bits

Our main goal is to make the two error bounds for erasures and quantization comparable. To this end, we use systematic binary error-correcting codes for packets of quantized coefficients, and replace a portion of the output from the sigma-delta quantizer by the check bits.

Definition. A binary (n, k)-code is an invertible map

$$C : \mathbb{Z}_2^k \to \mathbb{Z}_2^n\,.$$

The *minimum distance* of this code is the minimal number of bits by which any two code words (elements in the range of C) differ. A *systematic (n, k)-code* simply appends check bits, meaning $q = (q_1, q_2, \dots q_k)$ maps to $C(q) = (q'_1, q'_2, \dots q'_n)$ such that $q'_j = q_j$ for all $j \in \{1, 2, \dots k\}$.

The relevance of this definition is that among any block of n transmitted bits, the minimum distance is the number of bit erasures that cannot be corrected any more.

The reconstruction strategy we study is given by incorporating check bits in the output of the quantizer, which are used by the receiver to correct a portion of the erased bits. The remaining, incorrectible bits are then omitted from reconstruction.

As already mentioned, we will exploit a particular accompanying quantization strategy, which we briefly explain.

Definition. Let Q be the binary mid-riser quantizer with stepsize $\delta > 0$ and let $\mathcal{F} = \{f_1, f_2, \dots f_N\}$ be an N/d-tight frame for a d-dimensional real Hilbert space $\mathcal{H}$. Also, assume that C is a binary (n, k)-code. Given an input vector $x \in \mathcal{H}$, then the *C-embedded sigma-delta quantization of x* is $Q_{\mathcal{F},C}(x) = \frac{d}{N}\sum_{j=1}^{N} q_j f_j$, where the sequence $\{q_j\}_{j=1}^{\infty}$ associated with the initialization value $u_0 = 0$ is defined by

$$q_{m+j} := \begin{cases} Q(\langle x, f_{m+j}\rangle + u_{m+j-1}), & 1 \leq j \leq k\,, \\ C((q_{m+1}, q_{m+2}, \dots q_{m+k}))_j, & \text{else}\,, \end{cases}$$

for any $m \in \{0, n, 2n, \dots\}$, and $j \in \{1, 2, \dots n\}$, and the map for updating the internal variable is

$$u_{m+j} := \langle x, f_{m+j}\rangle - q_{m+j} + u_{m+j-1}\,.$$

Our first main theorem is the stability of this modified sigma-delta algorithm.

Theorem. Let Q be a binary mid-riser quantizer with stepsize $\delta > 0$, let $\mathcal{F} = \{f_1, f_2, \dots f_N\}$ be an N/d-tight equal-norm frame for a d-dimensional real Hilbert space $\mathcal{H}$, and let C be a systematic binary (n, k)-code, such that n divides N. If $\|x\| \leq \alpha\delta/2$, $\alpha < 1$, and

$$k \geq \frac{n}{2}(1 + \alpha)$$

then in the course of the C-embedded first-order sigma-delta quantization, the internal variable is bounded by

$$|u_j| \leq \frac{\delta}{2}((n - k + 1) + (n - k)\alpha) \leq \delta(k - \frac{k^2}{n} + \frac{1}{2})$$

for all $j \in \{1, 2, \dots N\}$.

Proof. We proceed by induction. At the end of the first block of n bits, if all $n - k$ check bits were chosen incorrectly and the input is taken to be the worst case, then u_N reaches the maximum magnitude stated in the theorem. In the course of quantizing the next block, due to the bound on the input, each bit allows the quantizer to recover at

least $\delta/2 - \alpha\delta/2$. With the inequality $k \geq \frac{n}{2}(1+\alpha)$ we deduce

$$k(\frac{1}{2} - \frac{\alpha}{2}) \geq \frac{1}{2}(n-k)(1+\alpha)$$

which means u_j is contained in $[-\delta/2, \delta/2]$ before the next check bit is encountered. □

Similarly as in [1] and [3], we deduce an error estimate from the bound on the internal variable.

The relevant quantity in this estimate is derived from the frame geometry, as in [3],

$$T(\mathcal{F}) = \|(f_1 - f_2) \pm (f_2 - f_3) \pm \cdots \pm (f_{N-1} - f_N) \pm f_N\| .$$

We define the maximal error caused by quantization to be

$$eq(V, \delta, \alpha) = \max_{\|x\| \leq \alpha\delta/2} \|Q_{\mathcal{F},C}(x) - x\| ,$$

where V is the analysis operator of the frame $\mathcal{F}$.

Theorem. Under the same assumptions as in the preceding theorem,

$$eq(V, \delta, \alpha) \leq \frac{d}{N}(\frac{\delta}{2}((n-k)(1+\alpha)+1)T(\mathcal{F}) .$$

Proof. This is an immediate consequence of the bound on the internal variable and the proof in [3]. □

In comparison with the unmodified first order sigma-delta quantization, we have a bound that is worse by at most a factor of $2(n-k)$. However, the advantage of the embedded check bits is the ability to correct erasures in each block.

Assume the initial probability measure applies an erasure with a probability of p to each coefficient. Assume that the code C has minimal distance $np + t$ with $t > 0$. Let $\mathbb{P}'$ denote the probability measure governing the erasures remaining after the error correction has been applied in each block of length n.

Definition. The combination of quantization, erasures and error correction gives the reconstruction error

$$ec(V, \delta, \alpha, \mathbb{P}') = \mathbb{E}[\max_{\|x\| \leq \alpha\delta/2} \|\frac{1}{A}\sum_j \omega_j q_j f_j - x\|] ,$$

where $\omega_j = 0$ means that the j-th coefficient is erased.

The following lemma helps bound the probability of erasures remaining, if the weight of the code is larger than the expected number of erasures before correction.

Lemma. (Hoeffding [8]). Let $\mathbb{E}[\omega_j] = 1 - p$ and assume that the minimum distance of C is bounded below by $n(p+\epsilon)$, $\epsilon > 0$. The probability $p'$of an individual coefficient being erased after the error correction is applied is bounded by

$$p' \leq \exp(-2n\epsilon^2) .$$

Now we can combine the two error estimates for quantization and erasures.

Theorem. Let $\epsilon > 0$, assume C has minimal distance $n(p+\epsilon)$. Let $\mathbb{P}'$ be the probability measure governing the erasures after the error correction has been applied. Under the additional assumptions of the preceding theorem,

$$ec(V, \delta, \alpha, \mathbb{P}') \leq eq(V, C, \delta, \alpha) + \frac{d\delta}{2}\exp(-2n\epsilon^2) .$$

Proof. First we apply Minkowski's inequality to separate the error caused by quantization and by erasures. The expected number of erasures is Np', with p' bounded in accordance with the preceding lemma. Each erased coefficient has magnitude $\delta/2$, so the norm of the vectors which are omitted in the reconstruction can at most be $\delta d p'/2$. □

The remaining question is which asymptotics can be achieved for the minimum distance with a suitable sequence of codes.

To this end, we quote a version of the Gilbert-Varshamov bound.

Lemma. Let $0 \leq q \leq 1/2$, then there exist infinitely many systematic linear (n,k)-codes with minimum distance at least nq and rate

$$\frac{k}{n} \geq 1 - H_2(q) ,$$

where $H_2(q) = -q\log_2 q - (1-q)\log_2(1-q)$ is the binary entropy.

Proof. The usual form of the Gilbert Varshamov bound for linear codes [11, Ch. 17] can be re-stated as a bound for the maximal number of erasures that can be corrected by certain codes. In this form, it states the existence of linear codes for which any $n-d+1$ rows of the generator matrix have rank k if $d \geq nq$, meaning up to $d-1$ erasures can be corrected. Permuting the rows so that the first k have maximal rank and right-multiplying by the inverse of this $k \times k$ block gives the generator matrix for a systematic code that can correct the same number of erasures. □

We are ready to state the final result.

Theorem. Let $0 \leq p < q \leq 1/2$, $H_2(q) \leq (1-\alpha)/2$, $0 < \alpha < 1$ and denote $\epsilon = q - p$. Consider the sequence of systematic linear codes provided by the Gilbert-Varshamov bound for minimum distance bounded below by nq and let $N \geq ne^{2n\epsilon^2}$, then

$$ec(V, \delta, \alpha, \mathbb{P}') \leq \frac{d\delta}{2N}((2\ln N H_2(q)/\epsilon^2 + 1)T(\mathcal{F}) + \frac{1}{2\epsilon^2}\ln N) .$$

Proof. From the assumption, we have $e^{2n\epsilon^2} \leq N$ and thus $n \leq \frac{1}{2\epsilon^2}\ln N$. By the Gilbert-Varshamov bound,

$$n - k \leq nH_2(q) \leq \frac{1}{2\epsilon^2}\ln N H_2(q) .$$

Using the Hoeffding inequality on the error due to the remaining erasures gives

$$e^{-2n\epsilon^2} \leq \frac{n}{N} \leq \frac{1}{2\epsilon^2}\frac{\ln N}{N} .$$

Thus, the two error terms have the same asymptotic behavior. □

We note that this error bound is only worse by a term logarithmic in N compared to the quantization error without erasures. We also remark that even in the lossy regime, when the error correction fails with near certainty in any packet, then we still have $p' \leq p$ and thus

$$ec(V, \delta, \alpha, \mathbb{P}') \leq d\delta((\frac{\ln N}{N}H_2(q)/\epsilon^2 + 1)T(\mathcal{F}) + \frac{p}{2}) .$$

Acknowledgment

This work was partially supported by NSF DMS 07-04216, NSF DMS 08-07399 and by the Deutsche Forschungsgemeinschaft (DFG) under Heisenberg Fellowship KU 1446/8-1.

References:

[1] J. J. Benedetto, A. M. Powell, and O. Yilmaz, Sigma-Delta quantization and finite frames, *IEEE Trans. Inform. Theory* 52:1990–2005, 2006.

[2] B. G. Bodmann and V. I. Paulsen. Frames, graphs and erasures. *Linear Algebra Appl.* 404:118–146, 2005.

[3] B. G. Bodmann and V. I. Paulsen. Frame Paths and Error Bounds for Sigma-Delta Quantization. *Appl. Comput. Harmon. Anal.* 22:176–197, 2007.

[4] P. Boufounos, A. V. Oppenheim, and V. K. Goyal. Causal Compensation for Erasures in Frame Representations. *IEEE Trans. Signal Proc.* 3:1071–1082, 2008.

[5] P. Casazza and J. Kovačević, Equal-norm tight frames with erasures. (English summary) Frames. *Adv. Comput. Math.* 18:387–430, 2003.

[6] V. K. Goyal, J. Kovačević, and J. A. Kelner. Quantized frame expansions with erasures. *Appl. Comp. Harm. Anal.* 10:203–233, 2001.

[7] V. K. Goyal, J. Kovačević, and M. Vetterli. Quantized frame expansions as source-channel codes for erasure channels. In: Proc. Data Compr. Conf., Snowbird, UT, Mar. 1999.

[8] W. Hoeffding, Probability inequalities for sums of bounded random variables, *J. Amer. Stat. Assoc.* 58 (301):13–30, 1963.

[9] R. B. Holmes and V. I. Paulsen, Optimal frames for erasures, *Linear Algebra Appl.* 377:31–51, 2004.

[10] D. Kalra, Complex equiangular cyclic frames and erasures, *Linear Algebra Appl.* 419:373–399, 2006.

[11] F. J. MacWilliams and N. J. A. Sloane, The theory of error-correcting codes. North-Holland, Amsterdam, 1977

[12] D. Petz, Spectral scale of self-adjoint operators and trace inequalities, *J. Math. Anal. Appl.* 109:74–82, 1985.

[13] M. Püschel and J. Kovačević, Real, Tight Frames with Maximal Robustness to Erasures, *Proc. Data Compr. Conf.*, Snowbird, UT, March 2005, pp. 63–72.

[14] T. Strohmer and R. Heath, Grassmannian frames with applications to coding and communication, *Appl. Comput. Harmon. Anal.* 14:257–275, 2003.

Special session on

Efficient Design and Implementation of Sampling Rate Conversion, Resampling and Signal Reconstruction Methods

Chair: Hakan Johansson and Christian Vogel

Structures for Interpolation, Decimation, and Nonuniform Sampling Based on Newton's Interpolation Formula

Vesa Lehtinen and Markku Renfors

Department of Communications Engineering, Tampere University of Technology
P.O.Box 553, FI-33101 Tampere, Finland
{vesa.lehtinen,markku.renfors}@tut.fi

Abstract:

The variable fractional-delay (FD) filter structure by Tassart and Depalle performs Lagrange interpolation in an efficient way. We point out that this structure directly corresponds to Newton's interpolation (backward difference) formula, hence we prefer to refer to it as the *Newton FD filter*. This structure does not function correctly when the fractional delay is made time-variant, e.g., in sample rate conversion. We present a simple modification that enables time-variant usage such as fractional sample rate conversion and nonuniform resampling. We refer to the new structure as the *Newton (interpolator) structure*. Almost all advantages of the Newton FD structure are preserved. Furthermore, we suggest that by transposing the Newton interpolator we obtain the *transposed Newton structure* which can be used in decimation as well as reconstruction of nonuniformly sampled signals, analogously to the transposed Farrow structure. The presented structures are a competitive alternative for the Farrow structure family when low complexity and flexibility are required.

1. Introduction

In [1][2][3], Tassart and Depalle as well as Candan derive an efficient implementation structure for FD filters, depicted in Fig. 1, from Lagrange's interpolation formula. It turns out that the obtained filter structure directly corresponds to Newton's (backward difference) interpolation formula [4] (with some subexpression sharing) which indeed is equivalent with Lagrange interpolation [5]. Newton's backward difference formula is

$$f(t+\tau) = \sum_{m=0}^{\infty} \frac{\tau^{(m)} \Delta^m f(t)}{m!}, \tag{1}$$

where

$$\tau^{(m)} = \prod_{k=0}^{m-1} (\tau + k) \tag{2}$$

is the rising factorial, and Δ is the backward difference operator such that $\Delta^m f(t) = \Delta^{m-1} f(t) - \Delta^{m-1} f(t-1)$ and $\Delta^0 f(t) = f(t)$, resulting in

$$\Delta^m f(t) = \sum_{k=0}^{m} (-1)^k \binom{m}{k} f(t-k). \tag{3}$$

Newton's backward difference formula provides an efficient means to realise piecewise-polynomial interpolation for DSP. Its complexity is only $O(M)$ (where M is the interpolator order)–cf. equivalent Lagrange implementations based on the Farrow structure [6] having $O(M^2)$ complexity [3]. The subfilters are multiplier-free and extremely simple. The structure is modular, as highlighted by the grey shading in Fig. 1, and the interpolator order can be changed in real time [3].

Unfortunately, the structure presented in Fig. 1 does not function correctly in sample rate conversion (SRC). Because the multiplications are performed between the subfilters, making them time-variant will result in incorrect output. This is because each output sample should only depend on the current value of the delay parameter D; in Fig. 1, past values of D contribute to the output through the delayed paths through the subfilters. Therefore, the structure in Fig. 1 is only useful in single-rate, time-invariant or slowly-varying fractional-delay filtering.

We propose a slightly modified structure that allows arbitrary resampling, including increasing the sample rate by arbitrary, also fractional, factors (fractional interpolation). We also point out that the structure can be transposed to obtain a decimator structure that possesses all the advantages of the Newton interpolation structure.

This work was supported by the Graduate School in Electronics, Telecommunications and Automation (GETA).

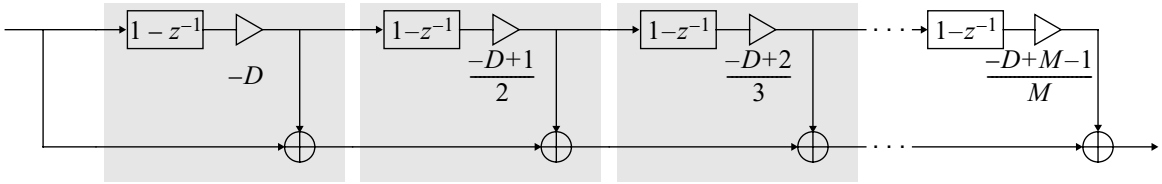


Figure 1. The fractional-delay filter structure proposed in [1][3], based on Newton's interpolation formula.

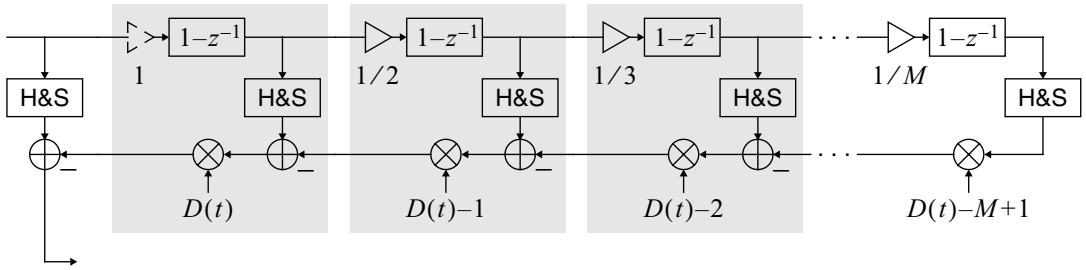


Figure 2. The Newton interpolator structure suitable for sample rate conversion. The hold & sample (H&S) blocks perform the sampling at the output sample instants.

2. The Newton structure for interpolation

In order to allow fractional SRC and arbitrary resampling, the Newton structure must work correctly with a time-variant fractional delay. This is achieved through two simple steps: (*i*) We invert the summation order at the output part of the structure from that presented in [1][3] (this was already done in [2]). (*ii*) The time-varying multiplications can now be implemented in the high-rate part between the adders. The improved structure is shown in Fig. 2. We refer to it as the *Newton interpolator structure* or the Newton structure for short. Also the improved structure is modular, permitting changing the interpolator order in real time. In single-rate FD filtering, the improved structure is equivalent to [1][2][3].

In Fig. 2, the H&S blocks stand for hold & sample, i.e., each output sample obtains the value of the previously arrived input sample.

In fractional interpolation, i.e., increasing the sample rate by a fractional factor, we use the common notation illustrated in Fig. 3. The time interval between the previous input sample and the next output sample to be generated is expressed using the *fractional interval* variable μ which is normalised with respect to the input sample interval so that $\mu \in [0, 1)$.

Interpolation of uniformly spaced input samples can be modelled as convolution [5], leading to the generic model depicted in Fig. 4 [7]. The continuous-time (CT) linear time-invariant (LTI) model filter is piecewise polynomial, with $M + 1$ pieces, each with duration equal to the input sample interval T_{in}. Hence the impulse response length is $(M + 1)T_{\text{in}}$.

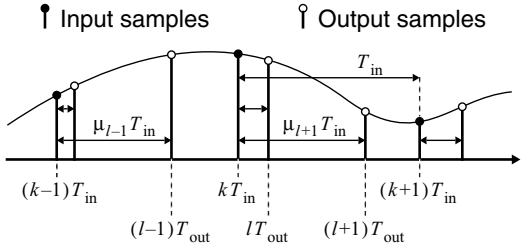


Figure 3. Definition of the fractional interval μ for interpolation.

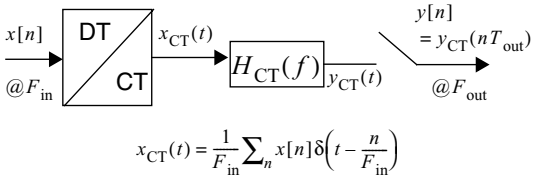


Figure 4. The generic model for SRC by arbitrary factors.

The composite transfer function of m cascaded subfilters is

$$(1 - z^{-1})^m = \sum_{n=0}^{m} (-1)^n \binom{m}{n} z^{-n}, \tag{4}$$

cf. (3). The output of the interpolator is

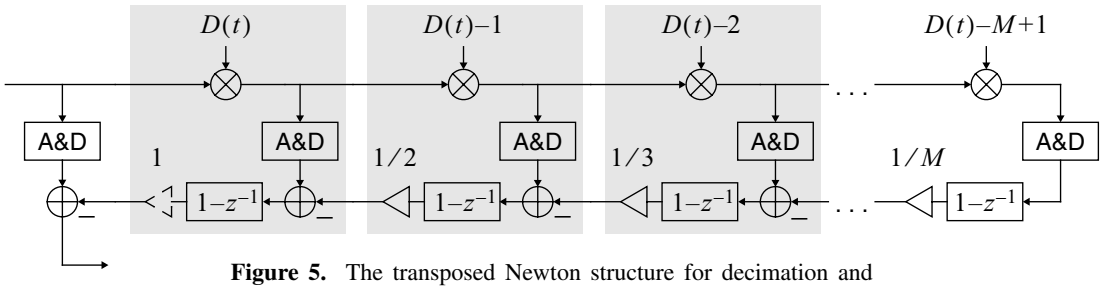


Figure 5. The transposed Newton structure for decimation and reconstruction of signals from nonuniformly spaced samples.

$$y((k+\mu)T_{\text{in}}) = \sum_{n=0}^{M} h((n+\mu)T_{\text{in}})x[k-n]$$

$$= \sum_{n=0}^{M} x[k-n](-1)^n \sum_{m=n}^{M} (-1)^m \binom{m}{n} \frac{(D_0-\mu)_m}{m!} \quad (2.1)$$

where

$$\binom{m}{n} = 0, \quad n < 0 \vee n > m \quad (5)$$

for $m \geq 0$, and

$$(x)_m = \prod_{k=0}^{m-1} (x-k) \quad (6)$$

is the falling factorial. The delay of the interpolator is $D_0 T_{\text{in}}$. The parameter D_0 can be chosen quite freely, but the best amplitude response and linear phase response are obtained with $D_0 = (M+1)/2$ [1].

The continuous-time model impulse response of the interpolator is then (cf. the expression of the filter input in Fig. 4)

$$h((n+\mu)T_{\text{in}}) = \frac{1}{T_{\text{in}}} \sum_{m=n}^{M} (-1)^{n+m} \binom{m}{n} \frac{(D_0-\mu)_m}{m!}. \quad (7)$$

The reversed summation order in the high-rate part comes with a price: the structure is more costly to pipeline than those in [1][3] because the signal paths cannot share pipeline registers.

3. The transposed Newton structure

There exists a duality[1] between decimation and interpolation that allows transforming a decimator into an interpolator and vice versa through network transposition [7]. By transposing the Newton interpolator, we obtain the structure depicted in Fig. 5. We refer to this as the *transposed Newton structure*. The transpose is obtained by inverting the flow direction of all signals and replacing each block with its dual. For instance, the H&S block is replaced with the accumulate & dump (A&D) block, which sums up all its input samples since the previous output sample. This is also the most straightforward way to obtain the transposed Farrow structure from the Farrow structure[2] [9].

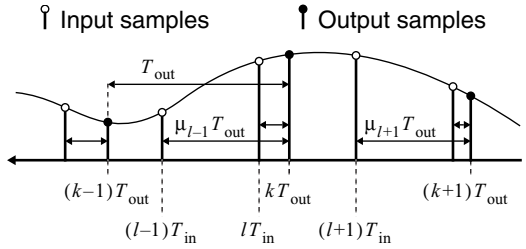


Figure 6. Definition of the fractional interval μ for the transposed structure (dual of interpolation).

The output samples of the transposed Newton structure are uniformly spaced, but the input samples may arrive at arbitrary time instants. The generic SRC model (Fig. 4) is valid also for the transposed Newton structure. The model impulse response is again piecewise-polynomial, now with the piece duration equal to the *output* sample interval. The model impulse response is obtained by replacing T_{in} with T_{out} in (7) and redefining μ according to Fig. 6 (reflecting the duality between decimation and interpolation). For an input sample arriving at time instant t, the fractional interval is

$$\mu(t) = \left\lceil \frac{t}{T_{\text{out}}} \right\rceil - \frac{t}{T_{\text{out}}} \in [0, 1). \quad (8)$$

For fractional decimation, the fractional interval for the l^{th} input sample is

$$\mu_l = \left\lceil \frac{lT_{\text{in}}}{T_{\text{out}}} \right\rceil - \frac{lT_{\text{in}}}{T_{\text{out}}}. \quad (9)$$

The impulse response in the generic model is now

$$h((n+\mu)T_{\text{out}}) = \frac{1}{T_{\text{out}}} \sum_{m=n}^{M} (-1)^{n+m} \binom{m}{n} \frac{(D_0-\mu)_m}{m!} \quad (10)$$

1. There exist a number of definitions for duality, including the adjoint. Here we use the generalised duality/transpose as defined in [7].

2. The structure in [8] (transposed structure I in [9]) is not the true transpose of the Farrow structure even though the duality of responses holds.

with integer n. Again, $D_0 = (M + 1)/2$ for the best response. In the frequency response, the model filter has $M + 1$ zeros at each (nonzero) integer multiple of the output sample rate, hence realising antialiasing regardless of the decimation factor.

The transposed Newton structure is able to receive input samples at arbitrary time instants, which makes it a potential building block for reconstruction of signals from nonuniformly spaced samples (e.g., in algorithms like [10][11]), as earlier suggested for the transposed Farrow structure in [12].

The transposed Newton structure shares the advantages and disadvantages of the Newton interpolator, such as modularity, $O(M)$ complexity and the inefficient zero locations.

4. Computational complexity

In interpolation by factor R, the Newton structure will perform $(1 + R)M$ additions and $(1 + R)M$ multiplications per input sample on average. In decimation by R, the transposed Newton structure will perform $(R - 1)(1 + M) + 2M$ additions and $(1 + R)M$ multiplications per output sample. The first term in the addition count comes from the A&D block. Multiplication by a constant inverse of a small integer requires only few additions/subtractions.

Unambiguous complexity comparison between the proposed structures and alternatives, mainly the Farrow family, would require specifying the implementation technology and the SRC factor. However, the following points can be made: (*i*) The basis multipliers are more complex in the Newton structures (integer part present in the time-variant coefficients) than in Farrow structures (no integer part). Hence, large SRC factors are unfavourable to the Newton family. (*ii*) If the Lagrange response suffices, the ultimate simplicity of the subfilters makes the Newton family superior to the Farrow structure when the SRC factor is small. (*iii*) The response of the Newton structures can be improved only by increasing the order (i.e., number of stages). In designs with a low oversampling factor and/or strict performance requirements, this may lead to a very high filter order. In such cases, an optimised Farrow design with a non-Lagrange response will have a lower complexity and smaller delay.

5. Conclusions

The proposed structures allow efficient piecewise Newton interpolation for SRC and arbitrary resampling as well as its dual for decimation and reconstruction of nonuniformly sampled signals. The advantages of the proposed structures include low, $O(M)$ complexity (high orders are feasible at the cost of a long delay), very simple subfilters and run-time adjustability of the filter order. As a drawback, the basis multipliers running at the high-rate end of the filter have longer wordlengths than in the Farrow counterparts.

Due to their simplicity, the Newton structures may be useful as building blocks of more complicated algorithms for interpolation, decimation, and reconstruction of nonuniformly sampled signals.

References:

[1] S. Tassart and Ph. Depalle, “Fractional delays using Lagrange interpolators,“ in *Proc. Nordic Acoustic Meeting*, Helsinki, Finland, 12–14 June, 1996.

[2] S. Tassart, Ph. Depalle, “Analytical approximations of fractional delays: Lagrange interpolators and allpass filters,” in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.* (*ICASSP*’97), 21–24 Apr 1997, pp. 455–458.

[3] Ç. Candan, “An efficient filtering structure for Lagrange interpolation,” *IEEE Signal Processing Letters*, Vol. 14, No. 1, Jan 2007, pp. 17–19.

[4] E.W. Weisstein, "Newton’s Backward Difference Formula." Available: http://mathworld.wolfram.com/NewtonsBackwardDifferenceFormula.html. Visited: 22 Jan 2008.

[5] E. Meijering, “A chronology of interpolation: From ancient astronomy to modern signal and image processing,” in *Proc. of the IEEE*, Vol. 90, No. 3, Mar 2002, pp. 319–342.

[6] C.W. Farrow, “A continuously variable digital delay element,” in *Proc. IEEE Int. Symp. Circ. Syst. (ISCAS’88)*, Espoo, Finland, June 1988, pp. 2641–2645.

[7] R.E. Crochiere, L.R. Rabiner, *Multirate Digital Signal Processing*, Prentice-Hall, 1983.

[8] T. Hentschel, G. Fettweis, “Continuous-time digital filters for sample-rate conversion in reconfigurable radio terminals,” in *Proc. European Wireless*, Dresden, Germany, Sep 2000, pp. 55–59.

[9] D. Babic, J. Vesma, T. Saramäki, M. Renfors, “Implementation of the transposed Farrow structure,” in *Proc. IEEE Int. Symp. Circ. Syst*., May 2002, pp. IV-5–IV-8.

[10] F. Marvasti, M. Analoui, M. Gamshadzahi, “Recovery of signals from nonuniform samples using iterative methods,” *IEEE Trans. Signal Proc*., Vol. 39, No. 4, Apr 1991, pp. 872–878.

[11] F.A. Marvasti, P.M. Clarkson, M.V. Dokic, U. Goenchanart, C. Liu, “Reconstruction of speech signals with lost samples,” *IEEE Trans. Signal Proc.*, Vol. 40, No. 12, Dec 1992, pp. 2897–2903.

[12] D. Babic and M. Renfors, “Reconstruction of non-uniformly sampled signal using transposed Farrow structure,” in *Proc. Int. Symp. Circ. Syst.* (*ISCAS*), Vancouver, Canada, May 2004, Vol. III, pp. 221–224.

Chromatic Derivatives, Chromatic Expansions and Associated Function Spaces

Aleksandar Ignjatović

School of Computer Science and Engineering, University of New South Wales,
and National ICT Australia (NICTA), Sydney, Australia;
ignjat@cse.unsw.edu.au

Abstract:

We present the basic properties of the chromatic derivatives and the chromatic expansions as well as a motivation for introducing these notions. The chromatic derivatives are special, numerically robust linear differential operators; the chromatic expansions are the associated local expansions, which possess the best features of both the Taylor and the Nyquist expansions. This makes them potentially useful in fields involving sampled data, such as signal and image processing.

1. Motivation

The Nyquist–(Whittaker–Kotelnikov–Shannon) expansion $f(t) = \sum_{n=-\infty}^{\infty} f(n) \sin \pi(t-n)/\pi(t-n)$ of a π-band limited signal of finite energy $f(t) \in \mathbf{BL}(\pi)$ is of *global nature*, because it requires samples of the signal at integers of arbitrarily large absolute value. On the other hand, since signals from $\mathbf{BL}(\pi)$ are analytic functions, they can also be represented by the Taylor expansion, $f(t) = \sum_{n=0}^{\infty} f^{(n)}(0)\ t^n/n!$. Such expansion is of *local nature*, because the values of the derivatives $f^{(n)}(0)$ are determined by the values of the signal in an arbitrarily small neighborhood of zero.

While the Nyquist expansion has a central role in digital signal processing, the Taylor expansion is of very limited use there, for several reasons.

(1) Numerical evaluation of higher order derivatives of a signal from its samples is very noise sensitive; in general, one is cautioned against numerical differentiation of signals given by empirical samples.

(2) The Taylor expansion of a signal $f \in \mathbf{BL}(\pi)$ converges non-uniformly; its truncations are unbounded and have rapid error accumulation.

(3) The Nyquist expansion of a signal $f \in \mathbf{BL}(\pi)$ converges to f in $\mathbf{BL}(\pi)$ and thus the action of a filter A on any $f \in \mathbf{BL}(\pi)$ can be expressed using the samples of f and the impulse response $A[\mathrm{sinc}\,]$ of A, i.e.,

$$A[f](t) = \sum_{n=-\infty}^{\infty} f(n)\ A\,[\mathrm{sinc}\,]\,(t-n). \qquad (1)$$

In contrast, the polynomials obtained by truncating the Taylor series do not belong to $\mathbf{BL}(\pi)$ and nothing similar to (1) holds for the Taylor expansion.

The chromatic derivatives and the chromatic expansions and approximations were introduced to obtain local signal representations which do not suffer from these problems.

2. Chromatic Derivatives

To explain our notions, we first consider normalized and rescaled Legendre polynomials $P_n^L(\omega)$ which satisfy

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} P_n^L(\omega)\ P_m^L(\omega)\mathrm{d}\omega = \delta(m-n),$$

and then define operator polynomials

$$\mathcal{K}_t^n = \frac{1}{\mathrm{i}^n} P_n^L\left(\mathrm{i}\ \frac{d}{dt}\right). \qquad (2)$$

It is easy to verify that for $f \in \mathbf{BL}(\pi)$ and its Fourier transform $\widehat{f(\omega)}$ we have

$$\mathcal{K}^n[f](t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{i}^n P_n^L(\omega)\widehat{f(\omega)}\ \mathrm{e}^{\mathrm{i}\,\omega t} d\omega.$$

Figure 1 compares the plots of $P_n^L(\omega)$ and ω^n/π^n for $n = 15$ to $n = 18$, which are the transfer functions (save a factor of i^n) of the operators $\mathcal{K}^n$ and of the (normalized) derivatives $1/\pi^n\ \mathrm{d}^n/\mathrm{d}t^n$, respectively. While the transfer functions of the normalized "standard derivatives" $1/\pi^n\ \mathrm{d}^n/\mathrm{d}t^n$ obliterate the spectrum of the signal, leaving only its edges which in practice contain mostly noise, the transfer functions of operators $\mathcal{K}^n$ form a family of well separated, interleaved and increasingly refined comb filters. Due to their spectrum preserving property, we call the operators $\mathcal{K}^n$ *the chromatic derivatives* associated with the Legendre polynomials. Both analytic estimates and empirical tests have shown that the chromatic derivatives

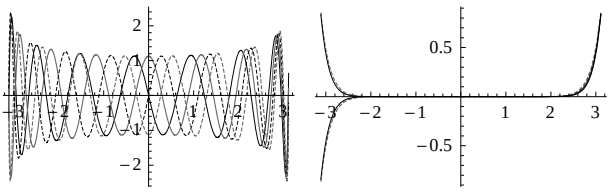


Figure 1: Graphs of $P_n^L(\omega)$ (left) and ω^n/π^n (right), for $n = 15 - 18$.

can be accurately and robustly evaluated from samples of the signal taken at a small multiple (2 to 4) of the usual Nyquist rate, thus solving problem (1) associated with numerical evaluation of the standard derivatives, mentioned above. Chromatic expansions, on the other hand, were introduced to solve problems (2) and (3).

3. Chromatic Approximations

Proposition 1 *Let $\mathcal{K}^n$ be the chromatic derivatives associated with the Legendre polynomials, and let $f(t)$ be any analytic function; then for all t,*

$$f(t) = \sum_{n=0}^{\infty} (-1)^n \; K^n[f](u) \; K^n[\mathrm{sinc}\,](t-u). \quad (3)$$

If $f \in \mathbf{BL}(\pi)$ the series converges uniformly and in L_2.

The series in (3) is denoted by $\mathrm{CE}[f,u](t)$ and is called *the chromatic expansion* of $f(t)$ associated with the Legendre polynomials; a truncation of this series up to first $n+1$ terms is denoted by $\mathrm{CA}[f,n,u](t)$ and is called a *chromatic approximation* of $f(t)$. Just like a Taylor approximation, a chromatic approximation is also a local approximation: its coefficients are the values of differential operators $\mathcal{K}^m[f](u)$ at a single instant u, and for all $k \leq n$, $f^{(k)}(u) = \mathrm{d}^k/\mathrm{d}t^k \; \mathrm{CA}[f,n,u](t)\big|_{t=u}$.
Figure 2 compares the behavior of the chromatic approximation (black) of a signal $f \in \mathbf{BL}(\pi)$ (gray) with the behavior of the Taylor approximation of $f(t)$ (dashed). Both approximations are of order sixteen. The plot reveals that, when approximating a signal $f \in \mathbf{BL}(\pi)$, a chromatic approximation has a much gentler error accumulation when moving away from the point of expansion than the Taylor approximation of the same order.
Functions $\mathcal{K}^n[\mathrm{sinc}\,](t)$ appearing in the chromatic expansion associated with the Legendre polynomials are given by $\mathcal{K}^n[\mathrm{sinc}\,](t) = (-1)^n\sqrt{2n+1}\;\; j_n(\pi t)$, where j_n is the spherical Bessel function of the first kind of order n. Thus, unlike the monomials that appear in the Taylor formula, functions $\mathcal{K}^n[\mathrm{sinc}\,](t)$ belong to $\mathbf{BL}(\pi)$ and satisfy $|\mathcal{K}^n[\mathrm{sinc}\,](t)| \leq 1$ for all $t \in \mathbb{R}$. Consequently, the chromatic approximations are bounded on $\mathbb{R}$ and belong to $\mathbf{BL}(\pi)$. Also, as Proposition 1 asserts, the chromatic approximation of a signal $f \in \mathbf{BL}(\pi)$ converges in $\mathbf{BL}(\pi)$. Thus, if A is a filter, then A commutes with the differential operators $\mathcal{K}^n$ and for every $f \in \mathbf{BL}(\pi)$, we have the following analogue of (1):

$$A[f](t) = \sum_{n=0}^{\infty} (-1)^n \; \mathcal{K}^n[f](0) \; \mathcal{K}^n[A[\,\mathrm{sinc}\,]](t).$$

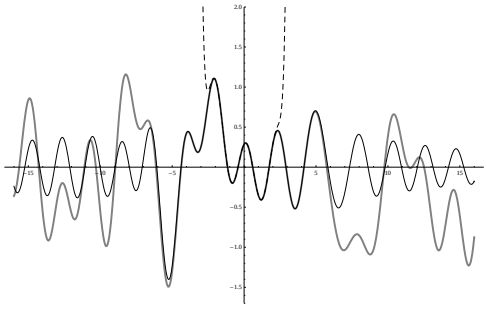


Figure 2: Chromatic approximation (black) and Taylor's approximation (dashed) of a signal from $\mathbf{BL}(\pi)$ (gray).

Thus, while local, the chromatic expansion possesses the features that make the Nyquist expansion useful in signal processing. This, together with numerical robustness of the chromatic derivatives, makes chromatic approximations applicable in fields involving empirically sampled data, such as digital signal and image processing.
The next proposition demonstrates another remarkable feature of the chromatic derivatives which is relevant to signal processing.

Proposition 2 *Let $\mathcal{K}^n$ be the chromatic derivatives associated with the (re-scaled and normalized) Legendre polynomials, and $f, g \in \mathbf{BL}(\pi)$. Then*

$$\begin{aligned}
\sum_{n=0}^{\infty} K^n[f](t)^2 &= \int_{-\infty}^{\infty} f(x)^2 dx; \\
\sum_{n=0}^{\infty} K^n[f](t) K^n[g](t) &= \int_{-\infty}^{\infty} f(x) g(x) dx; \\
\sum_{n=0}^{\infty} K^n[f](t) K_t^n[g(u-t)] &= \int_{-\infty}^{\infty} f(x) g(u-x) dx.
\end{aligned}$$

Thus, the sums on the left hand side of the above equations do not depend on the choice of the instant t.

Note that the above equations provide local representations of the usual norm, the scalar product and the convolution, respectively, which are defined in L_2 globally, as improper integrals.
Given the above properties of the Legendre polynomials, it is natural to ask if other families of orthonormal polynomials have similar properties. This question was answered in [1].

4. General Chromatic Derivatives

Let $\mathcal{M} : \mathcal{P}_\omega \to \mathbb{R}$ be a linear functional on the vector space $\mathcal{P}_\omega$ of real polynomials in the variable ω. Such $\mathcal{M}$ is called a moment functional and $\mu_n = \mathcal{M}(\omega^n)$ is the moment of $\mathcal{M}$ of order n.

Definition 1 *A moment functionals $\mathcal{M}$ is chromatic if it satisfies the following conditions (condition (iii) is not essential, but simplifies the technicalities):*
(i) $\mathcal{M}$ is positive definite;
(ii) $\limsup_{n\to\infty} \mu_n^{1/n}/n < \infty$;
(iii) $\mathcal{M}$ is symmetric, i.e., $\mu_{2n+1} = 0$ for all n.

For functionals $\mathcal{M}$ which satisfy conditions (i) and (iii) there exists a family of real polynomials $\{P_n^{\mathcal{M}}(\omega)\}_{n\in\mathbb{N}}$, such that $P_n^{\mathcal{M}}(\omega)$ contains only powers of ω of the same parity as n and which are orthonormal with respect to $\mathcal{M}$; i.e., for all m, n,

$$\mathcal{M}(P_m^{\mathcal{M}}(\omega)\, P_n^{\mathcal{M}}(\omega)) = \delta(m-n).$$

The family $\{P_n^{\mathcal{M}}(\omega)\}_{n\in\mathbb{N}}$ is a family of orthonormal polynomials which corresponds to a symmetric positive definite moment functional $\mathcal{M}$ just in case there exists a sequence of positive reals $\{\gamma_n\}_{n\in\mathbb{N}}$ such that

$$P_{n+1}^{\mathcal{M}}(\omega) = \tfrac{1}{\gamma_n}\,\omega\,P_n^{\mathcal{M}}(\omega) - \tfrac{\gamma_{n-1}}{\gamma_n}\,P_{n-1}^{\mathcal{M}}(\omega). \quad (4)$$

For every positive definite moment functional there exists a non-decreasing bounded function $a(\omega)$, called an m–distribution function, such that for the associated Stieltjes integral we have

$$\int_{-\infty}^{\infty}\omega^n\,\mathrm{d}a(\omega) = \mu_n, \quad (5)$$

$$\int_{-\infty}^{\infty}P_n^{\mathcal{M}}(\omega)\,P_m^{\mathcal{M}}(\omega)\,\mathrm{d}a(\omega) = \delta(m-n). \quad (6)$$

If $\mathcal{M}$ is chromatic, then condition (3) implies that $\{P_n^{\mathcal{M}}(\omega)\}_{n\in\mathbb{N}}$ is a complete system in $L^2_{a(\omega)}$.
Let $\varphi \in L^2_{a(\omega)}$; we can define a corresponding function $f_\varphi : \mathbb{R}\to\mathbb{C}$ by

$$f_\varphi(t) = \int_{-\infty}^{\infty}\varphi(\omega)e^{\mathrm{i}\omega t}\,\mathrm{d}a(\omega), \quad (7)$$

and one can show that (7) can be differentiated under the integral sign any number of times. Setting

$$\mathcal{K}^n = \tfrac{1}{\mathrm{i}^n}P_n^{\mathcal{M}}(\omega)\left(\mathrm{i}\,\tfrac{\mathrm{d}}{\mathrm{d}t}\right)$$

we get that for all t

$$\mathcal{K}^n[f_\varphi](t) = \int_{-\infty}^{\infty}\mathrm{i}^n\,P_n^{\mathcal{M}}(\omega)\,\varphi(\omega)\,\mathrm{e}^{\mathrm{i}\omega t}\mathrm{d}a(\omega), \quad (8)$$

i.e., $\langle\varphi(\omega)\mathrm{e}^{\mathrm{i}\omega t}, P_n^{\mathcal{M}}(\omega)\rangle_{a(\omega)} = (-\mathrm{i})^n\mathcal{K}^n[f_\varphi](t)$. Thus, $\varphi(\omega)\mathrm{e}^{\mathrm{i}\omega t} = (-\mathrm{i})^n\mathcal{K}^n[f_\varphi](t)P_n^{\mathcal{M}}(\omega)$, and by Parseval's Theorem, for every $t\in\mathbb{R}$,

$$\textstyle\sum_{n=0}^{\infty}|\mathcal{K}^n[f_\varphi](t)|^2 = \|\varphi(\omega)\mathrm{e}^{\mathrm{i}\omega t}\|^2_{a(\omega)} = \|\varphi(\omega)\|^2_{a(\omega)}.$$

Thus, if $\varphi\in L^2_{a(\omega)}$, then the sum $\sum_{n=0}^{\infty}|\mathcal{K}^n[f_\varphi](t)|^2$ converges to a constant function on $\mathbb{R}$.
If we let

$$\boldsymbol{m}(t) = \int_{-\infty}^{\infty}\mathrm{e}^{\mathrm{i}\omega t}\mathrm{d}a(\omega), \quad (9)$$

then (5) implies $\boldsymbol{m}^{(k)}(0) = \mathrm{i}^k\mu_k$. It can be shown that condition (iii) of Definition 1 implies that $\boldsymbol{m}(t)$ is analytic at every $t\in\mathbb{R}$ (moreover, it is analytic on a strip in $\mathbb{C}$; see [2]). For the chromatic approximation associated with $\mathcal{M}$,

$$\mathrm{CA}^{\mathcal{M}}[f,n,u](t) = \textstyle\sum_{k=0}^{n}(-1)^k\mathcal{K}^k[f](u)\mathcal{K}^k[\boldsymbol{m}](t-u),$$

one can show that

$$|f_\varphi(t) - \mathrm{CA}^{\mathcal{M}}[f_\varphi,n,u](t)| < \textstyle\sum_{k=n+1}^{\infty}\left|\mathcal{K}^k[f_\varphi](u)\right|^2.$$

Thus, $f_\varphi(t) = \sum_{k=0}^{\infty}(-1)^k\mathcal{K}^k[f_\varphi](u)\,\mathcal{K}^k[\boldsymbol{m}](t-u)$, and the convergence is uniform on $\mathbb{R}$.

Definition 2 *$L_2^{\mathcal{M}}$ denotes the space of functions analytic on $\mathbb{R}$ which satisfy $\sum_{k=0}^{\infty}\mathcal{K}^k[f](0)^2<\infty$.*

Let $f(t)\in L_2^{\mathcal{M}}$; then

$$\varphi_f(\omega) = \textstyle\sum_{k=0}^{\infty}(-\mathrm{i})^k\mathcal{K}^k[f](0)P_k^{\mathcal{M}}(\omega)$$

belongs to $L^2_{a(\omega)}$ and for all t,

$$f(t) = \int_{-\infty}^{\infty}\varphi_f(\omega)\,\mathrm{e}^{\mathrm{i}\omega t}\,\mathrm{d}a(\omega).$$

On the space $L_2^{\mathcal{M}}$ one can now introduce locally defined norm, inner product and convolution using equations from Proposition 2, and for every fixed u, the chromatic expansion of an $f\in L_2^{\mathcal{M}}$ is just the Fourier series of f in the orthonormal and complete base $\{\mathcal{K}_u^n[\boldsymbol{m}(t-u)]\}_{n\in\mathbb{N}}$.

5. Examples

Example 1. (Legendre polynomials/Spherical Bessel functions) Let $L_n(\omega)$ be the Legendre polynomials; if we set $P_n^L(\omega) = \sqrt{2n+1}\,L_n(\omega/\pi)$, then

$$\textstyle\int_{-\pi}^{\pi}P_n^L(\omega)P_m^L(\omega)\frac{\mathrm{d}\omega}{2\pi} = \delta(m-n).$$

The corresponding recursion coefficients in equation (4) are given by the formula $\gamma_n = \pi(n+1)/\sqrt{4(n+1)^2-1}$. In this case $\boldsymbol{m}(t) = \operatorname{sinc}t$, and $\mathcal{K}^n[\boldsymbol{m}](t) = (-1)^n\sqrt{(2n+1)}\;j_n(\pi t)$, where $j_n(x)$ is the spherical Bessel function of the first kind of order n. The corresponding space $L_2^{\mathcal{M}}$ consists of all analytic functions which belong to L_2 and have a Fourier Transform supported in $[-\pi,\pi]$.

Example 2. (Chebyshev polynomials of the first kind/Bessel functions) Let $P_n^T(\omega)$ be the family of orthonormal polynomials obtained by normalizing and rescaling the Chebyshev polynomials of the first kind, $T_n(\omega)$, by setting $P_0^T(\omega)=1$ and $P_n^T(\omega) = \sqrt{2}\,T_n(\omega/\pi)$ for $n>0$. In this case

$$\textstyle\int_{-\pi}^{\pi}P_n^T(\omega)P_m^T(\omega)\,\frac{\mathrm{d}\omega}{\pi^2\sqrt{1-\left(\frac{\omega}{\pi}\right)^2}} = \delta(n-m).$$

The corresponding function (9) is $\boldsymbol{m}(t) = J_0(\pi t)$ and $\mathcal{K}^n[\boldsymbol{m}](t) = (-1)^n\sqrt{2}\,J_n(\pi t)$ for $n>0$, where $J_n(t)$ is the Bessel function of the first kind of order n. In the recurrence relation (4) the coefficients are given by $\gamma_0 = \pi/\sqrt{2}$ and $\gamma_n = \pi/2$ for $n>0$. The corresponding space $L_2^{\mathcal{M}}$ consists of analytic functions whose Fourier transform $\widehat{f(\omega)}$ is supported in $(-\pi,\pi)$ and satisfies $\int_{-\pi}^{\pi}\sqrt{1-(\omega/\pi)^2}\;|\widehat{f(\omega)}|^2\mathrm{d}\omega<\infty$. The chromatic expansion of a function $f(t)$ is the Neumann series

$$f(t) = f(0)J_0(\pi t) + \sqrt{2}\;\textstyle\sum_{n=1}^{\infty}\mathcal{K}^n[f](0)J_n(\pi t).$$

Thus, the chromatic expansions corresponding to various families of orthogonal polynomials can be seen as generalizations of the Neumann series, while the families of corresponding functions $\{\mathcal{K}^n[\boldsymbol{m}](t)\}_{n\in\mathbb{N}}$ can be seen as generalizations (and a uniform representation) of some familiar families of special functions.

Example 3. (Hermite polynomials/Gaussian monomial functions) Let $H_n(\omega)$ be the Hermite polynomials; then the polynomials given by $P_n^H(\omega) = (2^n\,n!)^{-1/2}\,H_n(\omega)$ satisfy

$$\textstyle\int_{-\infty}^{\infty}P_n^H(\omega)P_m^H(\omega)\;\frac{\mathrm{e}^{-\omega^2}}{\sqrt{\pi}}\;\mathrm{d}\omega = \delta(n-m).$$

The corresponding function defined by (9) is $\boldsymbol{m}(t) = \mathrm{e}^{-t^2/4}$ and $\mathcal{K}^n[\boldsymbol{m}](t) = (-1)^n\,t^n\mathrm{e}^{-t^2/4}/\sqrt{2^n\,n!}$. The corresponding recursion coefficients are given by $\gamma_n = \sqrt{(n+1)/2}$. The corresponding space $L_2^{\mathcal{M}}$ consists of analytic functions whose Fourier transform $\widehat{f(\omega)}$ satisfies $\int_{-\infty}^{\infty}|\widehat{f(\omega)}|^2\;\mathrm{e}^{\omega^2}\mathrm{d}\omega<\infty$. The chromatic expansion of $f(t)$ is just the Taylor expansion of $f(t)\,\mathrm{e}^{t^2/4}$, multiplied by $\mathrm{e}^{-t^2/4}$.

6. Weakly Bounded Moment Functionals

To study local (i.e., non-uniform) convergence of chromatic expansions, we somewhat restrict the class of moment functionals we consider.

Definition 3 *Let $\mathcal{M}$ be a symmetric positive definite moment functional and let $\gamma_n > 0$ be such that* (4) *holds.*

(i) $\mathcal{M}$ is weakly bounded if there exist some $M \geq 1$, some $0 \leq p < 1$ and some integer r, such that for all $n \geq 0$, $1/M \leq \gamma_n \leq M(n+r)^p$ and $\gamma_n/\gamma_{n+1} \leq M^2$.

(ii) $\mathcal{M}$ is bounded if there exists some $M \geq 1$ such that $1/M \leq \gamma_n \leq M$ for all $n \geq 0$.

Thus, every bounded moment functional is also weakly bounded with $p = 0$. Functionals in our Example 1 and Example 2 are bounded. For bounded moment functionals the corresponding m-distribution $a(\omega)$ has a finite support and consequently $\boldsymbol{m}(t)$ is a band-limited signal. However, $\boldsymbol{m}(t)$ can be of infinite energy (i.e., not in L_2) as is the case in our Example 2. Moment functional in Example 3 is weakly bounded but not bounded ($p = 1/2$). We note that all important examples of classical orthogonal polynomials which correspond to weakly bounded moment functionals in fact satisfy a stronger condition $0 < \lim_{n\to\infty} \gamma_n/n^p < \infty$ for some $0 \leq p < 1$.

Lemma 3 *If $\mathcal{M}$ is a weakly bounded moment functional, then $\lim_{k\to\infty} (\mu_k/k!)^{1/k} = 0$. Thus, $\mathcal{M}$ is chromatic; moreover, $\boldsymbol{m}(z) = \sum_{n=0}^{\infty} \mathrm{i}^n \mu_n z^n/n!$ is an entire function on $\mathbb{C}$.*

Lemma 4 *Let $\mathcal{M}$ be weakly bounded and $p < 1$ as in Definition 3(i); then for every integer $k \geq 1/(1-p)$ there exists $K > 0$ and a polynomial $P(x)$ such that for every $n \in \mathbb{N}$ and every $z \in \mathbb{C}$,*

$$|\mathcal{K}^n[\boldsymbol{m}](z)| < |Kz|^n P(|z|) \mathrm{e}^{|Kz|^k}/n!^{1-p}.$$

This Lemma is used to prove the following Proposition.

Proposition 5 *Let $\mathcal{M}$ be as in Lemma 4, $f(z)$ an entire function and $u \in \mathbb{C}$. If $\lim_{n\to\infty} |f^{(n)}(u)/n!^{1-p}|^{1/n} = 0$, then the chromatic expansion of $f(z)$ centered at u converges everywhere to $f(z)$, and the convergence is uniform on every disc of finite radius.*

Thus, if $\mathcal{M}$ is bounded ($p = 0$) and f is an entire function, then the chromatic expansion $CE[f, u](t)$ converges to $f(t)$ for all t.

Many well known equalities for the Bessel functions $J_n(t)$ are just the special cases of chromatic expansions. For example, the chromatic expansions of $f(t) = \mathrm{e}^{\mathrm{i}\omega t}$, $f(t) = 1$ and $f(t) = \boldsymbol{m}(t+u)$ yield

$$\mathrm{e}^{\mathrm{i}\omega t} = \sum_{n=0}^{\infty} \mathrm{i}^n P_n^{\mathcal{M}}(\omega)\, \mathcal{K}^n[\boldsymbol{m}](t);$$

$$\boldsymbol{m}(t) + \sum_{n=1}^{\infty} \left(\prod_{k=1}^{n} \frac{\gamma_{2k-2}}{\gamma_{2k-1}} \right) \mathcal{K}^{2n}[\boldsymbol{m}](t) = 1,$$

$$\boldsymbol{m}(t+u) = \sum_{n=0}^{\infty} (-1)^n \mathcal{K}^n[\boldsymbol{m}](u) \mathcal{K}^n[\boldsymbol{m}](t),$$

which generalize the following well known equalities:

$$\mathrm{e}^{\mathrm{i}\,\omega t} = J_0(t) + 2\sum_{n=1}^{\infty} \mathrm{i}^n T_n(\omega) J_n(t);$$

$$J_0(t) + 2\sum_{n=1}^{\infty} J_{2n}(t) = 1;$$

$$J_0(t+u) = J_0(u)J_0(t) + 2\sum_{n=1}^{\infty} (-1)^n J_n(u) J_n(t).$$

7. Non-Separable Inner Product Spaces

If $\mathcal{M}$ is weakly bounded, the periodic functions do not belong to $L_2^{\mathcal{M}}$; for example, $\sum_{n=0}^{\infty} \mathcal{K}^n[\sin \omega t]^2$ diverges. We now consider some inner product spaces in which pure harmonic oscillations have finite positive norms ([3, 2]).

Definition 4 *Assume again that $\mathcal{M}$ is weakly bounded and let p be as in Definition 3. We denote by $\mathcal{C}^{\mathcal{M}}$ the vector space of analytic functions such that the sequence*

$$\nu_n^f(t) = 1/(n+1)^{1-p} \sum_{k=0}^{n} \mathcal{K}^k[f](t)^2$$

converges uniformly on every finite interval.

Proposition 6 *Let $f, g \in \mathcal{C}^{\mathcal{M}}$ and*

$$\sigma_n^{fg}(t) = 1/(n+1)^{1-p} \sum_{k=0}^{n} \mathcal{K}^k[f](t) \mathcal{K}^k[g](t);$$

then the sequence $\{\sigma_n^{fg}(t)\}_{n\in\mathbb{N}}$ converges to a constant function. In particular, $\nu_n^f(t)$ is constant.

Corollary 7 *Let $\mathcal{C}_0^{\mathcal{M}}$ be the vector space consisting of analytic functions $f(t)$ such that $\lim_{n\to\infty} \nu_n^f(t) = 0$; then in the quotient space $\mathcal{C}_2^{\mathcal{M}} = \mathcal{C}^{\mathcal{M}}/\mathcal{C}_0^{\mathcal{M}}$ the limit $\lim_{n\to\infty} \sigma_n^{fg}(t)$ is independent of t and defines a scalar product on $\mathcal{C}_2^{\mathcal{M}}$.*

Proposition 8 *Let $\mathcal{M}$ correspond to Chebyshev polynomials as in our Example 2; then functions $f_\omega(t) = \sqrt{2}\sin \omega t$ and $g_\omega(t) = \sqrt{2}\cos \omega t$ for all $0 < \omega < \pi$ form an uncountable orthonormal system of vectors in $\mathcal{C}_2^{\mathcal{M}}$.*

Proposition 9 *Let $\mathcal{M}$ correspond to Hermite polynomials as in our Example 3; then for all $\omega > 0$ functions $f_\omega(t) = \sin \omega t$ and $g_\omega(t) = \cos \omega t$ form an uncountable orthogonal system of vectors in $\mathcal{C}_2^{\mathcal{M}}$, and $\|f_\omega\|^{\mathcal{M}} = \|g_\omega\|^{\mathcal{M}} = \mathrm{e}^{\omega^2/2}/\sqrt[4]{2\pi}$.*

Conjecture 1 *Assume that for some $0 \leq p < 1$ the recursion coefficients γ_n in* (4) *are such that γ_n/n^p converges to a finite positive limit. Then, for the corresponding family of orthogonal polynomials we have*

$$0 < \lim_{n\to\infty} 1/(n+1)^{1-p} \sum_{k=0}^{n} P_k^{\mathcal{M}}(\omega)^2 < \infty$$

for all ω in the support $sp(a)$ of the corresponding m-distribution function $a(\omega)$. Thus, in the corresponding space $\mathcal{C}_2^{\mathcal{M}}$ all pure harmonic oscillations with positive frequencies $\omega \in sp(a)$ have finite positive norm and are mutually orthogonal.

Detailed presentation of the theory of chromatic derivatives can be found in our references; preprints of some unpublished manuscripts are available at `http://www.cse.unsw.edu.au/~ignjat/diff`.

References:

[1] A. Ignjatovic. Local approximations based on orthogonal differential operators. *Journal of Fourier Analysis and Applications*, 13(3), 2007.

[2] A. Ignjatovic. Chromatic derivatives and associated function spaces. *manuscript*, 2008.

[3] A. Ignjatovic. Chromatic derivatives and local approximations. *to appear in: IEEE Transactions on Signal Processing*, 2009.

Estimation of the Length and the Polynomial Order of Polynomial-based Filters

Djordje Babic[(1)], and Heinz G. Göckler[(2)]

(1) Faculty of Computer Science, University Union, Belgrade, Knez Mihailova 6/VI, 11000 Belgrade, Serbia.
(2) DISPO, Faculty of Electrical Engineering and Information Sciences, Ruhr-Universität, Bochum, Germany.
djbabic@raf.edu.rs, goeckler@nt.rub.de

Abstract:

In many signal processing applications it is beneficial to use polynomial-based interpolation filters for sampling rate conversion. Actual implementations of these filters can be performed effectively by using the Farrow structure or its modifications. In the literature, several design methods have been proposed. However, estimation formulae for the number of polynomial-segments defining the finite length of the underlying continuous-time filter impulse response and the order of polynomials have not been known. This contribution presents estimation formulae for the length and the polynomial order of polynomial-based filters for various types of requirements. The formulae presented here can save time in designing, since they provide good starting values of length and order for a given set of requirements.

1. Introduction

In many signal processing applications it is required to determine signal samples at arbitrary positions between existing samples of a discrete-time signal. In these cases, it is beneficial to use polynomial-based interpolation filters. For these filters, an efficient overall implementation can be achieved by using a continuous-time impulse response $h_a(t)$ having the following properties [1], [2]; First, $h_a(t)$ is nonzero only in a finite interval $0 \le t < NT$ with N being an integer. Second, in each subinterval $nT \le t < (n+1)T$, for $n = 0, \ldots, N-1$, $h_a(t)$ is expressible as a polynomial of t of a given (low) order M. Third, $h_a(t)$ is symmetric with respect to $t = NT/2$ to guarantee phase linearity of the resulting overall system. The length of polynomial segments, T, can be selected to be equal to the input T_{in} or output T_{out} sampling interval, a fraction of the input or output sampling interval, or an integer multiple of the input or output sampling interval. The advantage of the above system lies in the fact that the actual implementation can be efficiently performed by using the Farrow structure [3] or its modifications [4], [5].

In the literature, several design methods have been proposed [1], [2], [4]. However, estimation formulae for the number N of polynomial-segments and the order M of polynomial have not been known. This contribution presents the missing estimation formulae for the length N and polynomial order M for various types of requirements. The formulae presented subsequently can save time for the filter designers, because they get suitable starting values for N and M that can be used for the given set of requirements. The formulae can also be used to estimate implementation costs of Farrow filter as subsystem of general sampling rate converters, for example, in optimal factorization of multistage decimation (interpolation).

2. Polynomial-based filters

As it has been originally suggested in [1], [2] when deriving the modified Farrow structure for interpolation, it is beneficial to construct $h_a(t)$ as follows:

$$h_a(t) = \sum_{n=0}^{N-1} \sum_{m=0}^{M} c_m(n) f_m(n, T, t) \qquad (1)$$

where the number of polynomial segments N is an integer. The basis functions $f_m(n, T, t)$, as defined in [1], are given by

$$f_m(n,T,t) = \begin{cases} \left(\dfrac{2(t-nT)}{T} - 1 \right)^m & \text{for } nT \le t < (n+1)T \\ 0 & \text{otherwise,} \end{cases} \qquad (2)$$

where the common polynomial order of all segments is M. The coefficients $c_m(n)$ are the adjustable parameters being related to each other by

$$c_m(N-1-n) = \begin{cases} c_m(n) & \text{for } m \text{ even} \\ -c_m(n) & \text{for } m \text{ odd} \end{cases} \qquad (3)$$

for $n = 0, 1, \ldots, N-1$, as consequence of the symmetry properties required above. The resulting $h_a(t)$ is characterized by the following properties: (i) $h_a(t)$ is nonzero for $0 \le t < NT$ and zero elsewhere; (ii) in each subinterval $nT \le t < (n+1)T$ for $n = 0, \ldots, N-1$, $h_a(t)$ is expressed as a polynomial of degree M; (iii) $h_a(t)$ is symmetric about $t = NT/2$, that is, $h_a(NT-t) = h_a(t)$.

Based on Property (iii), it is guaranteed that the resulting overall system has a linear phase, a very attractive property for many applications. Furthermore, the generation of the above $h_a(t)$ guarantees that, in the frequency domain, the zero-phase frequency response, when omitting the linear-phase term, is expressible as (see [1] for details)

$$H_a(j2\pi f) = \sum_{n=0}^{N/2-1} \sum_{m=0}^{M} c_m(n) G_m(n, T, f), \qquad (4)$$

where $G_m(n, T, f)$ is the Fourier transform of

$$g_m(n,T,t) = (-1)^m f_m(n,T,t-NT/2) + f_m(N-1-n,T,t-NT/2). \quad (5)$$

Since the above approximating function is linear with respect to the unknown coefficients $c_m(n)$, it enables one to optimize the overall filter to meet the given criteria in a manner similar to that used for synthesizing various types of linear-phase FIR filters [6]. In the above, T, the length of the polynomial segments, can be used to define different implementation structures as discussed in [4], [5]. As seen in [4], [5], T can be chosen as $T=\beta T_{in}$ or $T=\beta T_{out}$, where β is unity, an integer, or one divided by an integer. The selection depends on whether decimation or interpolation is under consideration, and on the structural needs for efficient implementation. The actual implementation can be efficiently performed by using the Farrow structure [3] or its modifications [4], [5].

For all these structure the number of fixed coefficients depends on the number N of polynomial segments and the order M of the polynomial in each segment. The total number of multipliers, exploiting the symmetry properties of (3), is given by

$$S = \begin{cases} N\cdot(M+1)/2 & \text{for } N \text{ even} \\ (N-1)(M+1)/2 + \lceil (M+1)/2 \rceil & \text{for } N \text{ odd.} \end{cases}, \quad (6)$$

For the purpose of illustration, the modified Farrow structure [1] is used with $T=T_{in}$. It should be pointed out that, in a practical realization, the coefficients' symmetry of the FIR branches will be exploited, and a single delay line can be shared with all branches.

3. Review of minimax design method

This section reviews minimax design method of polynomial-based filters of arbitrary length and order, as presented in [1], [2], for which we estimate N and M.

To this end, we assume a lowpass signal $x(n)\leftrightarrow X(e^{j\Omega in})$. Its sampling rate $F_{in}=1/T_{in}$ shall be converted by an arbitrary ration according to $F_{out}=RF_{in}$ yielding $y(l)\leftrightarrow Y(e^{j\Omega out})$. In case of $R>1$ ($R<1$) the system realizes interpolation (decimation). The ultimate aim is to determine a continuous-time, finite-length impulse response $h_a(t)$ of the sampling rate conversion system such that the Fourier transform of $h_a(t)$ meets following requirements [4] , [7]:

$$\begin{aligned} &(1-\delta_p) \le H_a(f) \le (1+\delta_p) && \text{for } |f| \le f_p = \alpha F/2 \\ &|H_a(f)| \le \delta_s && \text{for } |f| \in \Phi_s, \end{aligned} \quad (7)$$

where

$$\Phi_s = \begin{cases} [F/2, \infty] & \text{for Case A} \\ \bigcup_{k=1}^{\infty} [kF - f_p, kF + f_p] & \text{for Case B} \\ [F - f_p, \infty] & \text{for Case C.} \end{cases} \quad (8)$$

In all three cases, the signal is preserved according to the given tolerance in the passband region [0, f_p]. Furthermore, the aliasing components are attenuated in the defined manner. In Case A, all components aliasing into the baseband [0, $F/2$] are attenuated. In Case B, all components aliasing into the passband [0, f_p] are attenuated, but aliasing is allowed in the transition band [f_p, $F/2$]. In Case C, aliasing into the transition band [f_p, $F/2$] is allowed only from the band [$F/2$, $F+f_p$]. In the above discussion and in (7) and (8) F stands for F_{out} in a decimation case, and F_{in} in an interpolation case.

The minimax optimization method introduced in [1], [2] is probably the most convenient and the most flexible solution for designing polynomial-based interpolation filters:

Minimax Optimization Problem: Given N, M, and a compact subset $\Phi \subset [0,\infty)$ as well as a desired function $D(f)$ being continuous for $f \in \Phi$ and a weight function $W(f)$ being positive for $f \in \Phi$, find the $(M+1)N/2$ unknown coefficients $c_m(n)$ to minimize

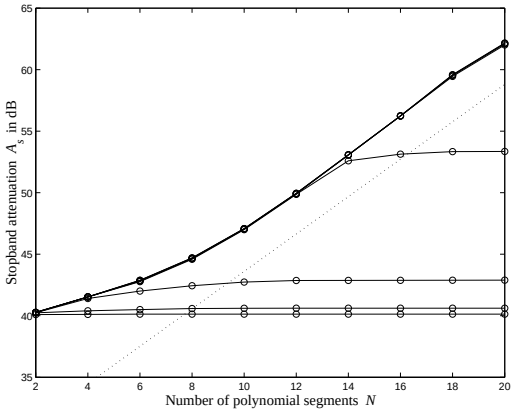
$$\delta_\infty = \max_{f\in\Phi} |W(f)[H_a(f) - D(f)]| \quad (9)$$

subject to the given time-domain conditions of $h_a(t)$. Here, $H_a(f)$ is the real-valued frequency response and $D(f)$ is the desired function according to specifications. (For details refer to [2]). The design procedure has been generalized, and modified for optimization of prolonged and transposed prolonged polynomial-based filters [4].

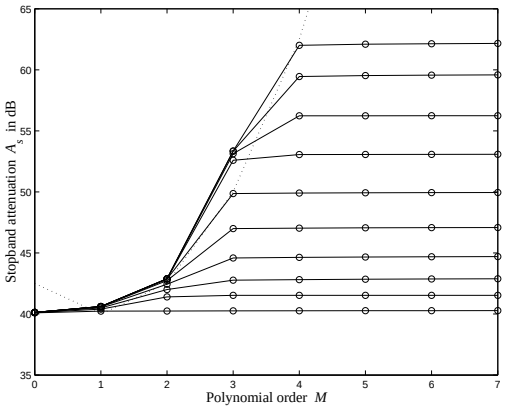
The minimax design method has several design parameters. First of all, the design parameters include passband and stopband regions Φ_p and Φ_s. The desired filter may have several passbands and stopbands as stated in [2]. Next, the minimum stopband attenuation δ_s, and maximum allowable passband ripple δ_p are also included. Other design parameters are the number of polynomial segments N and the order M of the polynomial, which determine the number of multipliers in the overall structure, see (6). Finally, some weighting function can be used to give different weights to passband and stopband [2]. Hence we give estimation formulae for the number N of polynomial segments and the order M of polynomial for a minimax design.

4. Estimation of *N* and *M*

In the previous section, we have seen that the number of polynomial segments N and the order M of the polynomial, are the design parameters that highly influence the performance of the filter in the frequency domain. Furthermore, the cost of realization, i.e. the number of multipliers, of a filter can be estimated by introducing the required values for N and M into (6). It would be very beneficial to estimate N and M by only using the given specifications of the filter in the frequency domain. Similar order estimation formulae exist for FIR filters, for example Kaiser order estimation [6], [8]. In the actual implementation, polynomial-based filters can be modeled as FIR filters [4]. Thus, we can start from the Kaiser formula and adapt it to polynomial-based filters. To this end, a lot of filters were designed, by using different system specifications, in order to adapt the Kaiser formula to polynomial-based case. The obtained estimation formula for the number of polynomial segments N, is rather similar to Kaiser formula for the order estimation of FIR filters. The



(a)



(b)

Fig. 1. *Case A specifications*: The passband and stopband edges are at f_p=0.4F_{in} and at f_s=0.5F_{in}, and stopband weighting W=100. (a) The curves are shown for M equals 0 to 7. Dashed line is plot obtained from the estimation formula for N. (b) The curves are shown for N equals 2 to 20. Dashed line is plot obtained from the estimation formula for M.

estimation formula for N, which can be found in [9], is not accurate enough. Hence, we propose the more accurate formula:

$$N_e = 2\left\lceil \frac{-20\log_{10}(\sqrt{\delta_p \delta_s}) - 8.4}{30.4(f_s - f_p)/F} \right\rceil \qquad (10)$$

where δ_p and δ_s are the maximum deviations of the amplitude response from unity for $f \in [0, f_p]$ and the maximum deviation from zero for $f \in \Phi_s$, respectively. Here, $\lceil x \rceil$ stands for the smallest integer which is larger or equal to x. It has been observed that in most cases the above estimation formula is rather accurate with only a 2% error. The formula above is valid for all three types of requirements, i.e., A, B, and C, as given by (7) and (8). However, if the transition band is narrow, i.e., in the case when $(f_s-f_p)/F \leq 0.1$, the required value of N should be increased by 2. Further, in the case of very narrow transition band ($(f_s-f_p)/F \leq 0.05$) the formula can not be used.

The kernel of the estimation formula for the number N of polynomial segments can be expressed in a different form:

$$N_e = 2\left\lceil \frac{A_s - 10\log_{10}(W) - 8.4}{30.4(f_s - f_p)/F} \right\rceil \qquad (11)$$

where A_s=-20log$_{10}$(δ_s) is the required attenuation in stopband, and $W=\delta_p/\delta_s$ represents weighting between required tolerances in passband and stopband.

The next problem is to find the minimum value of the polynomial order M to meet the specifications. It has been observed that the required value of M depends on the type of requirements from (7) and (8). Never the less, it is possible to consider the following estimate as good starting point for all three types of requirements:

$$M_e = \left\lceil \sqrt{\frac{A_s - 20 \cdot \log_{10}(W)}{2.5}} + \log_{10}(W) \right\rceil + 1. \qquad (12)$$

It has been observed that if transition band is relatively large to the sampling frequency, that is when $(f_s-f_p)/F \geq 0.5$, the required value of polynomial order M is lowered by one. The estimation formula cannot be used when the transition band is very small, i.e., in the case when $(f_s-f_p)/F<0.1$. However, even in this border situation required value of M is always smaller than M_e given by (12). Thus, the estimation formula (12) for the polynomial order M can be used to estimate the upper border for M for all types of requirements.

5. Design Examples

This part gives several examples to illustrate the performance of the presented formulae. To illustrate this, the following specifications are considered:

Case A *specifications*: The passband and stopband edges are at f_p=0.4F_{in} and at f_s=0.5F_{in}.

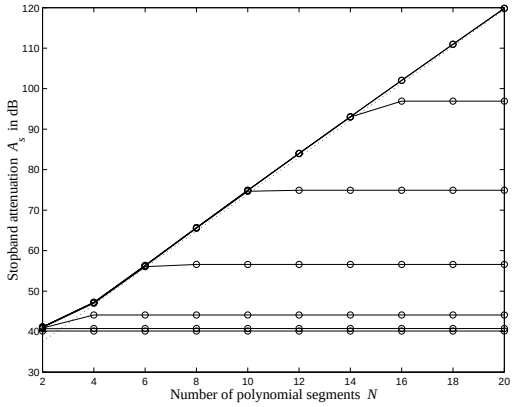
Case B *specifications*: The passband and stopband edges are at f_p=0.35F_{in} and at f_s=0.65F_{in}.

Case C *specifications*: The passband and stopband edges are at f_p=0.35F_{in} and at f_s=0.65F_{in}.

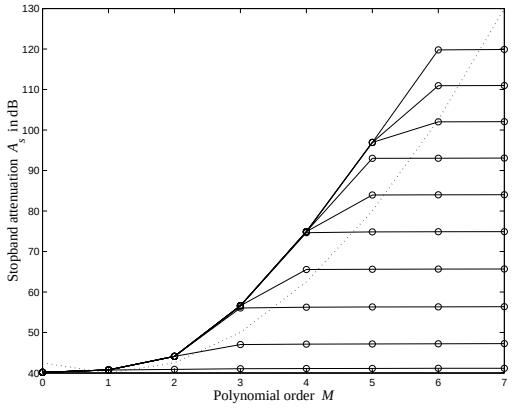
In each case, several filters have been designed in minimax sense with the passband weighting equal to unity and stopband weightings of W=100. The degree of the polynomial in each subinterval M varies from 0 to 7. The number of intervals N varies from 2 to 20. Recall that N is an even integer. Figures 1 give the results for Case A, the similar results for Case B are given in Fig. 2, and for Case C in Fig. 3. It can be observed that the estimation formulae are relatively good, as they estimate the border performance for the given set of requirements (dashed lines in Figs 1-3).

6. Conclusions

In this paper, the estimation formulae for the number N of polynomial segments and the polynomial order M are presented. It has been shown that these estimates give the border performance of the filter for the given set of specifications. Formulae for N and M can be used to estimate the starting value of these two parameters in minimax optimization. Furthermore, the formulae for N and M can be used to estimate implementation costs of

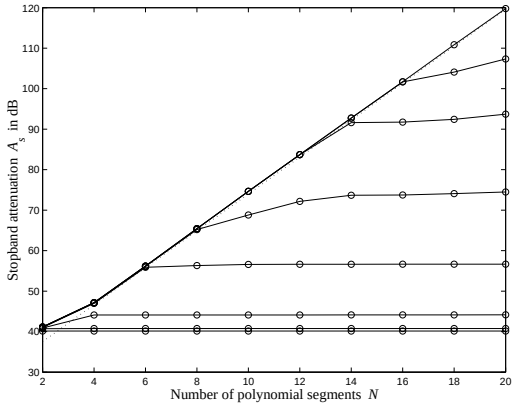


(a)

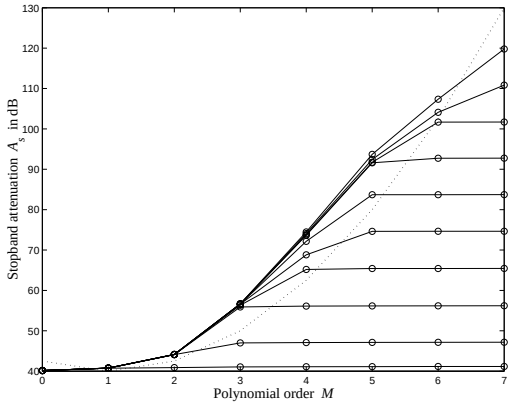


(b)

Fig. 2. *Case* B *specifications*: The passband and stopband edges are at f_p=0.35F_{in} and at f_s=0.65F_{in}, and stopband weighting W=100. (a)The curves are shown for M equals 0 to 7. Dashed line is plot obtained from the estimation formula for N. (b) The curves are shown for N equals 2 to 20. Dashed line is plot obtained from the estimation formula for M.



(a)



(b)

Fig. 3. *Case C specifications*: The passband and stopband edges are at f_p=0.35F_{in} and at f_s=0.65F_{in}, and stopband weighting W=100. (a) The curves are shown for M equals 0 to 7. Dashed line is plot obtained from the estimation formula for N. (b) The curves are shown for N equals 2 to 20. Dashed line is plot obtained from the estimation formula for M.

the Farrow based filters for the given set of requirements. Formulae can also be used to estimate implementation costs of composed sampling rate converters containing Farrow, for example, in optimal factorization for multistage decimation (interpolation).

References:

[1] J. Vesma and T. Saramäki, "Interpolation filters with arbitrary frequency response for all-digital receivers," in *Proc. 1996 IEEE Int. Symp. Circuits and Systems*, Atlanta, Georgia, May 1996, pp. 568–571.

[2] J. Vesma and T. Saramäki, "Polynomial-based interpolation Filters - Part I: Filter synthesis," *Circuits, Systems, and Signal Processing*, vol. 26, no. 2, pp. 115-146, March/April 2007.

[3] C. W. Farrow, "A continuously variable digital delay element,"in *Proc. 1988 IEEE Int. Symp. Circuits and Systems*, Espoo, Finland, June 1988, pp. 2641–2645.

[4] D. Babic, T. Saramäki, M. Renfors, "Conversion between arbitrary sampling rates using polynomial-based interpolation filters," in *Proc. 2nd Int. TICSP Workshop on Spectral Methods and Multirate Signal Processing SMMSP'02* , Toulouse, France, September 2002, pp. 57–64.

[5] D. Babic, *Techniques for sampling rate conversion by arbitrary factors with applications in flexible communications receivers*, Doctoral Thesis, Tampere University of Technology, 2004.

[6] T. Saramäki, "Finite impulse response filter design," Chapter 4 in *Handbook for Digital Signal Processing*, edited by S. K. Mitra and J. F. Kaiser, John Wiley & Sons, New York, 1993.

[7] D. Babic, J. Vesma, T. Saramäki, M. Renfors, "Implementation of the transposed Farrow structure," in *Proc. 2002 IEEE Int. Symp. Circuits and Systems*, Scotsdale, Arizona, USA, 2002, vol. 4, pp. 4–8.

[8] J.F. Kaiser, "Nonrecursive Digital Filter Design Using the - sinh Window Function," *Proc. 1974 IEEE Symp. Circuits and Systems*, (April 1974), pp.20-23.

[9]T. Saramäki, "Multirate Signal Processing," *Lecture Notes*, http://www.cs.tut.fi/~ts/

Special session on

Geometric Multiscale Analysis

Chair: Gitta Kutyniok

The Continuous Shearlet Transform in Arbitrary Space Dimensions, Frame Construction, and Analysis of Singularities

S. Dahlke [1], G. Steidl [2] and G. Teschke [3]

(1) Philipps-Universität Marburg, FB12 Mathematik und Informatik, Hans-Meerwein Straße, Lahnberge, 35032 Marburg, Germany.
(2) Universität Mannheim, Fakultät für Mathematik und Informatik, Institut für Mathematik, 68131 Mannheim, Germany.
(3) University of Applied Sciences Neubrandenburg, Institute for Computational Mathematics in Science and Technology, Brodaer Str. 2, 17033 Neubrandenburg, Germany.

dahlke@mathematik.uni-marburg.de, steidl@math.uni-mannheim.de, teschke@hs-nb.de

Abstract:

This note is concerned with the generalization of the continuous shearlet transform to higher dimensions. Similar to the two-dimensional case, our approach is based on translations, anisotropic dilations and specific shear matrices. We show that the associated integral transform again originates from a square-integrable representation of a specific group, the full n-variate shearlet group. Moreover, we verify that by applying the coorbit theory, canonical scales of smoothness spaces and associated Banach frames can be derived. We also indicate how our transform can be used to characterize singularities in signals.

1. Introduction

Modern technology allows for easy creation, transmission and storage of huge amounts of data. Confronted with a flood of data, such as internet traffic, or audio and video applications, nowadays the key problem is to extract the relevant information from these sets. To this end, usually the first step is to decompose the signal with respect to suitable building blocks which are well–suited for the specific application and allow a fast and efficient extraction. In this context, one particular problem which is currently in the center of interest is the analysis of *directional* information. Due to the bias to the coordinate axes, classical approaches such as, e.g., wavelet or Gabor transforms are clearly not the best choices, and hence new building blocks have to be developed. In recent studies, several approaches have been suggested such as ridgelets [2], curvelets [3], contourlets [7], shearlets [14] and many others. For a general approach see also [13]. Among all these approaches, the shearlet transform stands out because it is related to group theory, i.e., this transform can be derived from a square-integrable representation $\pi : \mathcal{S} \to \mathcal{U}(L_2(\boldsymbol{R}^2))$ of a certain group $\mathcal{S}$, the so-called *shearlet group*, see [5]. Therefore, in the context of the shearlet transform, all the powerful tools of group representation theory can be exploited.

So far, the shearlet transform is well developed for problems in $\boldsymbol{R}^2$. However, for analyzing *higher-dimensional* data sets, there is clearly an urgent need for further generalizations and applications. This is exactly the concern of this paper. One particular field of application is the geometrical structure analysis of multi-dimensional data, e.g. multimodal spectral measurements in meteorology.
To our best knowledge, it seems that there exist only few results in this direction: some important progress has been achieved for the curvelet case in [1] and for surfacelets in [16]. However, for the shearlet approach the question has been completely open.

2. Multivariate Continuous Shearlet Transform

In this section, we introduce the shearlet transform on $L_2(\boldsymbol{R}^n)$. This requires the generalization of the two-dimensional parabolic dilation matrix and of the shear matrix, respectively. Let I_n denote the (n, n)-identity matrix and 0_n, resp. 1_n the vectors with n entries 0, resp. 1. For $a \in \boldsymbol{R}^* := \boldsymbol{R} \setminus \{0\}$ and $s \in \boldsymbol{R}^{n-1}$, we set

$$A_a := \begin{pmatrix} a & 0_{n-1}^{\mathrm{T}} \\ 0_{n-1} & \operatorname{sgn}(a)|a|^{\frac{1}{n}}\, I_{n-1} \end{pmatrix}$$

and

$$S_s := \begin{pmatrix} 1 & s^{\mathrm{T}} \\ 0_{n-1} & I_{n-1} \end{pmatrix}.$$

Lemma 1 *The set $\boldsymbol{R}^* \times \boldsymbol{R}^{n-1} \times \boldsymbol{R}^n$ endowed with the operation*

$$(a, s, t) \circ (a', s', t') = (aa', s + |a|^{1-1/n}\, s', t + S_s A_a t')$$

is a locally compact group $\mathcal{S}$ which we call full shearlet group*. The left and right Haar measures on $\mathcal{S}$ are given by*

$$d\mu_l(a, s, t) = \frac{1}{|a|^{n+1}}\, da\, ds\, dt$$

and

$$d\mu_r(a,s,t) = \frac{1}{|a|}\, da\, ds\, dt.$$

In the following, we use only the left Haar measure and use the abbreviation $d\mu = d\mu_l$. For $f \in L_2(\boldsymbol{R}^n)$ we define

$$\pi(a,s,t)f(x) = f_{a,s,t}(x) := |a|^{\frac{1}{2n}-1} f(A_a^{-1} S_s^{-1}(x-t)). \tag{1}$$

It is easy to check that $\pi : \mathcal{S} \to \mathcal{U}(L_2(\boldsymbol{R}^n))$ is a mapping from $\mathcal{S}$ into the group $\mathcal{U}(L_2(\boldsymbol{R}^n))$ of unitary operators on $L_2(\boldsymbol{R}^n)$. Recall that a *unitary representation* of a locally compact group G with the left Haar measure μ on a Hilbert space $\mathcal{H}$ is a homomorphism π from G into the group of unitary operators $\mathcal{U}(\mathcal{H})$ on $\mathcal{H}$ which is continuous with respect to the strong operator topology.

Lemma 2 *The mapping π defined by* (1) *is a unitary representation of $\mathcal{S}$.*

A nontrivial function $\psi \in L_2(\boldsymbol{R}^n)$ is called *admissible*, if

$$\int_{\mathcal{S}} |\langle \psi, \pi(a,s,t)\psi\rangle|^2 d\mu(a,s,t) < \infty.$$

If π is irreducible and there exits at least one admissible function $\psi \in L_2(\boldsymbol{R}^n)$, then π is called *square integrable*. The following result shows that the unitary representation π defined in (1) is square integrable.

Theorem 3 *A function $\psi \in L_2(\boldsymbol{R}^n)$ is admissible if and only if it fulfills the* admissibility condition

$$C_\psi := \int_{\boldsymbol{R}^n} \frac{|\hat{\psi}(\omega)|^2}{|\omega_1|^n}\, d\omega < \infty. \tag{2}$$

Then, for any $f \in L^2(\boldsymbol{R}^n)$, the following equality holds true:

$$\int_{\mathcal{S}} |\langle f, \psi_{a,s,t}\rangle|^2\, d\mu(a,s,t) = C_\psi \, \|f\|^2_{L_2(\boldsymbol{R}^n)}. \tag{3}$$

In particular, the unitary representation π is irreducible and hence square integrable.

An example of a continuous shearlet can be constructed as follows: Let ψ_1 be a continuous wavelet with $\hat{\psi}_1 \in C^\infty(\boldsymbol{R})$ and supp $\hat{\psi}_1 \subseteq [-2,-\frac{1}{2}] \cup [\frac{1}{2},2]$, and let ψ_2 be such that $\hat{\psi}_2 \in C^\infty(\boldsymbol{R}^{n-1})$ and supp $\hat{\psi}_2 \subseteq [-1,1]^{n-1}$. Then the function $\psi \in L^2(\boldsymbol{R}^n)$ defined by

$$\hat{\psi}(\omega) = \hat{\psi}(\omega_1, \tilde{\omega}) = \hat{\psi}_1(\omega_1)\, \hat{\psi}_2\left(\frac{1}{\omega_1}\,\tilde{\omega}\right)$$

is a continuous shearlet. The support of $\hat{\psi}$ is depiced for $\omega_1 \geq 0$ in Fig. 1.

3. Multivariate Shearlet Coorbit Theory

In this section we want to establish a coorbit theory based on the square integrable representation (1) of the shearlet group. We mainly follow the lines of [4]. For further information on coorbit space theory, the reader is referred to [8, 9, 10, 11, 12].

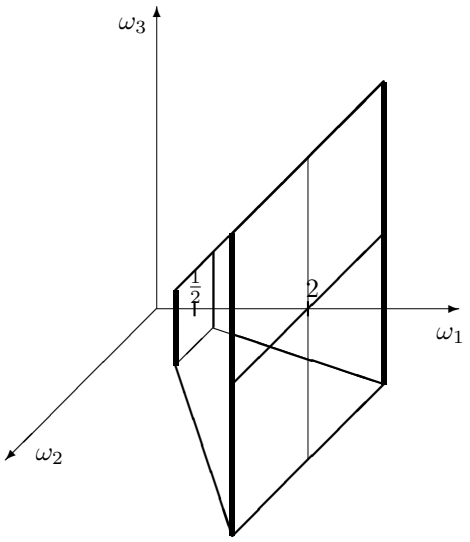


Figure 1: Support of the shearlet $\hat{\psi}$ for $\omega_1 \geq 0$.

3.1 Shearlet Coorbit Spaces

We consider weight functions $w(a,s,t) = w(a,s)$ that are locally integrable with respect to a and s, i.e., $w \in L_1^{loc}(\boldsymbol{R}^n)$ and fulfill $w((a,s,t)\circ(a',s',t')) \leq w(a,s,t)w(a',s',t')$ and $w(a,s,t) \geq 1$ for all $(a,s,t),(a',s',t') \in \mathcal{S}$. For $1 \leq p < \infty$, let

$$L_{p,w}(\mathcal{S}) := \{F \text{ measurable} :$$

$$\|F\|_{L_{p,w}(\mathcal{S})} := \left(\int_{\mathcal{S}} |F(g)|^p\, w(a,s,t)^p d\mu(a,s,t)\right)^{\frac{1}{p}} < \infty\},$$

and let $L_{\infty,w}$ be defined with the usual modifications. In order to construct the coorbit spaces related to the shearlet group we have to ensure that there exists a function $\psi \in L_2(\boldsymbol{R}^n)$ such that

$$\mathcal{SH}_\psi(\psi) = \langle \psi, \pi(a,s,t)\psi\rangle \in L_{1,w}(\mathcal{S}). \tag{4}$$

Fortunately, it turns out that (4) can be satisfied in our setting.

Theorem 4 *Let ψ be a Schwartz function such that* supp $\hat{\psi} \subseteq ([-a_1,-a_0] \cup [a_0,a_1]) \times Q_b$*,where* $Q_b := [-b_1,b_1] \times \cdots \times [-b_{n-1},b_{n-1}]$. *Then we have that $\mathcal{SH}_\psi(\psi) \in L_{1,w}(\mathcal{S})$, i.e.,*

$$\|\langle \psi, \pi(\cdot)\psi\rangle\|_{L_{1,w}(\mathcal{S})} =$$

$$\int_{\mathcal{S}} |\mathcal{SH}_\psi(\psi)(a,s,t)|\, w(a,s,t)\, d\mu(a,s,t) < \infty.$$

For ψ satisfying (4) we can consider the space

$$\mathcal{H}_{1,w} := \{f \in L_2(\boldsymbol{R}^n) : \mathcal{SH}_\psi(f) \in L_{1,w}(\mathcal{S})\}, \tag{5}$$

with norm $\|f\|_{\mathcal{H}_{1,w}} := \|\mathcal{SH}_\psi f\|_{L_{1,w}(\mathcal{S})}$ and its anti-dual $\mathcal{H}'_{1,w}$, the space of all continuous conjugate-linear functionals on $\mathcal{H}_{1,w}$. The spaces $\mathcal{H}_{1,w}$ and $\mathcal{H}'_{1,w}$ are π-invariant Banach spaces with continuous embeddings $\mathcal{H}_{1,w} \hookrightarrow \mathcal{H} \hookrightarrow \mathcal{H}'_{1,w}$, and their definition is independent of the shearlet ψ. Then the inner product on $L_2(\boldsymbol{R}^n) \times$

$L_2(\boldsymbol{R}^n)$ extends to a sesquilinear form on $\mathcal{H}^\sim_{1,w} \times \mathcal{H}_{1,w}$, therefore for $\psi \in \mathcal{H}_{1,w}$ and $f \in \mathcal{H}^\sim_{1,w}$ the *extended representation coefficients*

$$\mathcal{SH}_\psi(f)(a,s,t) := \langle f, \pi(a,s,t)\psi\rangle_{\mathcal{H}^\sim_{1,w}\times\mathcal{H}_{1,w}}$$

are well-defined. Now, for $1 \le p \le \infty$, we define the *shearlet coorbit spaces*

$$\mathcal{SC}_{p,w} := \{f \in \mathcal{H}^\sim_{1,w} : \mathcal{SH}_\psi(f) \in L_{p,w}(\mathcal{S})\} \quad (6)$$

with norms $\|f\|_{\mathcal{SC}_{p,w}} := \|\mathcal{SH}_\psi f\|_{L_{p,w}(\mathcal{S})}$. It holds that $\mathcal{SC}_{1,w} = \mathcal{H}_{1,w}$ and $\mathcal{SC}_{1,1} = L_2(\boldsymbol{R}^n)$.

3.2 Shearlet Banach Frames

The Feichtinger-Gröchenig theory provides us with a machinery to construct atomic decompositions and Banach frames for our shearlet coorbit spaces $\mathcal{SC}_{p,w}$. In a first step, we have to determine, for a compact neighborhood U of $e \in \mathcal{S}$ with non-void interior, so-called U–dense sets. A (countable) family $X = ((a,s,t)_\lambda)_{\lambda\in\Lambda}$ in $\mathcal{S}$ is said to be *U-dense* if $\cup_{\lambda\in\Lambda}(a,s,t)_\lambda U = \mathcal{S}$, and *separated* if for some compact neighborhood Q of e we have $(a_i,s_i,t_i)Q \cap (a_j,s_j,t_j)Q = \emptyset, i \ne j$, and *relatively separated* if X is a finite union of separated sets.

Lemma 5 *Let U be a neighborhood of the identity in $\mathcal{S}$, and let $\alpha > 1$ and $\beta,\gamma > 0$ be defined such that*

$$[\alpha^{\frac{1}{n}-1}, \alpha^{\frac{1}{n}}) \times [-\tfrac{\beta}{2},\tfrac{\beta}{2})^{n-1} \times [-\tfrac{\gamma}{2},\tfrac{\gamma}{2})^n \subseteq U. \quad (7)$$

Then the sequence

$$\{(\epsilon\alpha^j, \beta\alpha^{j(1-\frac{1}{n})}k, S_{\beta\alpha^{j(1-\frac{1}{n})}k} A_{\alpha^j}\gamma m) :$$

$$j \in \boldsymbol{Z}, k \in \boldsymbol{Z}^{n-1}, m \in \boldsymbol{Z}^n, \epsilon \in \{-1,1\}\} \quad (8)$$

is U-dense and relatively separated.

Next we define the *U–oscillation* as

$$\mathrm{osc}_U(a,s,t) :=$$

$$\sup_{u\in U} |\mathcal{SH}_\psi(\psi)(u\circ(a,s,t)) - \mathcal{SH}_\psi(\psi)(a,s,t)|. \quad (9)$$

Then, the following decomposition theorem, which was proved in a general setting in [8, 9, 10, 11, 12], says that discretizing the representation by means of an U-dense set produces an atomic decomposition for $\mathcal{SC}_{p,w}$.

Theorem 6 *Assume that the irreducible, unitary representation π is w-integrable and let an appropriately normalized $\psi \in L_2(\boldsymbol{R}^n)$ which fulfills*

$$M\langle\psi,\pi(a,s,t)\rangle := \sup_{u\in(a,s,t)U} |\langle\psi,\pi(u)\psi\rangle| \in L_{1,w}(\mathcal{S}) \quad (10)$$

be given. Choose a neighborhood U of e so small that

$$\|\,\mathrm{osc}_U\,\|_{L_{1,w}(\mathcal{S})} < 1. \quad (11)$$

Then for any U-dense and relatively separated set $X = ((a,s,t)_\lambda)_{\lambda\in\Lambda}$ the space $\mathcal{SC}_{p,w}$ has the following atomic decomposition: *If $f \in \mathcal{SC}_{p,w}$, then*

$$f = \sum_{\lambda\in\Lambda} c_\lambda(f)\pi((a,s,t)_\lambda)\psi \quad (12)$$

where the sequence of coefficients depends linearly on f and satisfies

$$\|(c_\lambda(f))_{\lambda\in\Lambda}\|_{\ell_{p,w}} \le C\|f\|_{\mathcal{SC}_{p,w}} \quad (13)$$

with a constant C depending only on ψ and with $\ell_{p,w}$ being defined by

$$\ell_{p,w} := \{c = (c_\lambda)_{\lambda\in\Lambda} : \|c\|_{\ell_{p,w}} := \|cw\|_{\ell_p} < \infty\},$$

where $w = (w((a,s,t)_\lambda))_{\lambda\in\Lambda}$. Conversely, if $(c_\lambda(f))_{\lambda\in\Lambda} \in \ell_{p,w}$, then $f = \sum_{\lambda\in\Lambda} c_\lambda\pi((a,s,t)_\lambda)\psi$ is in $\mathcal{SC}_{p,w}$ and

$$\|f\|_{\mathcal{SC}_{p,w}} \le C'\|(c_\lambda(f))_{\lambda\in\Lambda}\|_{\ell_{p,w}}. \quad (14)$$

Given such an atomic decomposition, the problem arises under which conditions a function f is completely determined by its *moments* $\langle f, \pi((a,s,t)_\lambda)\psi\rangle$ and how f can be reconstructed from these moments. This is answered by the following theorem which establishes the existence of Banach frames.

Theorem 7 *Impose the same assumptions as in Theorem 6. Choose a neighborhood U of e such that*

$$\|\,\mathrm{osc}_U\,\|_{L_{1,w}(\mathcal{S})} < 1/\|\mathcal{SH}_\psi(\psi)\|_{L_{1,w}(\mathcal{S})}. \quad (15)$$

Then, for every U-dense and relatively separated family $X = ((a,s,t)_\lambda)_{\lambda\in\Lambda}$ in G the set $\{\pi((a,s,t)_\lambda)\psi : \lambda \in \Lambda\}$ is a Banach frame for $\mathcal{SH}_{p,w}$. This means that

i) *$f \in \mathcal{SC}_{p,w}$ if and only if $(\langle f, \pi((a,s,t)_\lambda)\psi\rangle_{\mathcal{H}^\sim_{1,w}\times\mathcal{H}_{1,w}})_{\lambda\in\Lambda} \in \ell_{p,w}$;*

ii) *there exist two constants $0 < D \le D' < \infty$ such that*

$$D\,\|f\|_{\mathcal{SC}_{p,w}} \le \|(\langle f, \pi((a,s,t)_\lambda)\psi\rangle_{\mathcal{H}^\sim_{1,w}\times\mathcal{H}_{1,w}})_{\lambda\in\Lambda}\|_{\ell_{p,w}} \le D'\,\|f\|_{\mathcal{SC}_{p,w}}; \quad (16)$$

iii) *there exists a bounded, linear reconstruction operator $\mathcal{R}$ from $\ell_{p,w}$ to $\mathcal{SC}_{p,w}$ such that*
$\mathcal{R}\left((\langle f, \psi((a,s,t)_\lambda)\psi\rangle_{\mathcal{H}^\sim_{1,w}\times\mathcal{H}_{1,w}})_{\lambda\in\Lambda}\right) = f.$

It can be checked that the conditions (10), (11) and (15) can be satisfied, see [6] for details.

4. Analysis of Singularities

In this section, we deal with the decay of the shearlet transform at hyperplane singularities. The following analysis generalizes techniques and results presented in [15] for two dimensions. An $(n-m)$-dimensional hyperplane in $\boldsymbol{R}^n$, $1 \le m \le n-1$, not containing the x_1-axis can be written w.l.o.g. as

$$\underbrace{\begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}}_{x_A} + P \underbrace{\begin{pmatrix} x_{m+1} \\ \vdots \\ x_n \end{pmatrix}}_{x_E} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix},$$

$$P := \begin{pmatrix} p_1^{\mathrm{T}} \\ \vdots \\ p_m^{\mathrm{T}} \end{pmatrix} \in \boldsymbol{R}^{m,n-m}.$$

Then we obtain for

$$\nu_m := \delta(x_A + Px_E)$$

with the Delta distribution δ that

$$\begin{aligned} \hat{\nu}_m(\omega) &= \int_{\boldsymbol{R}^n} \delta(x_A + Px_E) e^{-2\pi i(\langle x_A,\omega_A\rangle + \langle x_E,\omega_E\rangle)}\, dx \\ &= \int_{\boldsymbol{R}^{n-m}} e^{-2\pi i(-\langle Px_E,\omega_A\rangle + \langle x_E,\omega_E\rangle)}\, dx_E \\ &= \delta(\omega_E - P^{\mathrm{T}}\omega_A). \end{aligned} \tag{17}$$

The following theorem describes the decay of the shearlet transform at hyperplane singularities. We use the notation $\mathcal{SH}_\psi f(a,s,t) \sim |a|^r$ as $a \to 0$, if there exist constants $0 < c \le C < \infty$ such that

$$c|a|^r \le \mathcal{SH}_\psi f(a,s,t) \le C|a|^r \quad \text{as } a \to 0.$$

Theorem 8 *Let $\psi \in L_2(\boldsymbol{R}^n)$ be a shearlet satisfying $\hat{\psi} \in C^\infty(\boldsymbol{R}^n)$. Assume further that $\hat{\psi}(\omega) = \hat{\psi}_1(\omega_1)\hat{\psi}_2(\tilde{\omega}/\omega_1)$, where $\operatorname{supp}\hat{\psi}_1 \in [-a_1,-a_0] \cup [a_0,a_1]$ for some $a_1 > a_0 \ge \alpha > 0$ and $\operatorname{supp}\hat{\psi}_2 \in Q_b$. If*

$$(s_m,\dots,s_{n-1}) = (-1,s_1,\dots,s_{m-1})\,P$$

and

$$(t_1,\dots,t_m) = -(t_{m+1},\dots,t_n)\,P^r,$$

then

$$\mathcal{SH}_\psi \nu_m(a,s,t) \sim |a|^{\frac{1-2m}{2n}} \quad \text{as } a \to 0. \tag{18}$$

Otherwise, the shearlet transform $\mathcal{SH}_\psi \nu_m$ decays rapidly as $a \to 0$.

Similar results can be derived for point singularities, see again [6] for details.

References:

[1] L. Borup and M. Nielsen, Frame decomposition of decomposition spaces, J. Fourier Anal. Appl., to appear.

[2] E. J. Candès and D. L. Donoho, *Ridgelets: a key to higher-dimensional intermittency?,* Phil. Trans. R. Soc. Lond. A. **357** (1999), 2495–2509.

[3] E. J. Candès and D. L. Donoho, *Curvelets - A surprisingly effective nonadaptive representation for objects with edges,* in Curves and Surfaces, L. L. Schumaker et al., eds., Vanderbilt University Press, Nashville, TN (1999).

[4] S. Dahlke, G. Kutyniok, G. Steidl, and G. Teschke, *Shearlet Coorbit Spaces and Associated Banach Frames*, Preprint Nr. 2007-5, Philipps-Universitt Marburg, 2007.

[5] S. Dahlke, G. Kutyniok, P. Maass, C. Sagiv, H.-G. Stark, and G. Teschke, *The uncertainty principle associated with the continuous shearlet transform*, Int. J. Wavelets Multiresolut. Inf. Process. **6** (2008), 157-181.

[6] S. Dahlke, G. Steidl, and G. Teschke, *The continuous shearlet transform in arbitrary space dimensions*, Preprint Nr. 2008–7, Philipps-Universität Marburg 2008.

[7] M. N. Do and M. Vetterli, *The contourlet transform: an efficient directional multiresolution image representation,* IEEE Transactions on Image Processing **14**(12) (2005), 2091–2106.

[8] H. G. Feichtinger and K. Gröchenig, *A unified approach to atomic decompositions via integrable group representations*, Proc. Conf. "Function Spaces and Applications", Lund 1986, Lecture Notes in Math. **1302** (1988), 52–73.

[9] H. G. Feichtinger and K. Gröchenig, *Banach spaces related to integrable group representations and their atomic decomposition I,* J. Funct. Anal. **86** (1989), 307–340.

[10] H. G. Feichtinger and K. Gröchenig, *Banach spaces related to integrable group representations and their atomic decomposition II,* Monatsh. Math. **108** (1989), 129–148.

[11] H. G. Feichtinger and K. Gröchenig, *Non–orthogonal wavelet and Gabor expansions and group representations,* in: Wavelets and Their Applications, M.B. Ruskai et.al. (eds.), Jones and Bartlett, Boston, 1992, 353–376.

[12] K. Gröchenig, *Describing functions: Atomic decompositions versus frames,* Monatsh. Math. **112** (1991), 1–42.

[13] K. Guo, W. Lim, D. Labate, G. Weiss, and E. Wilson, *Wavelets with composite dilations and their MRA properties.* Appl. Comput. Harmon. Anal. **20** (2006), 220–236.

[14] K. Guo, G. Kutyniok, and D. Labate, *Sparse multidimensional representations using anisotropic dilation und shear operators,* in Wavelets und Splines (Athens, GA, 2005), G. Chen und M. J. Lai, eds., Nashboro Press, Nashville, TN (2006), 189–201.

[15] G. Kutyniok and D. Labate, *Resolution of the wavefront set using continuous shearlets,* Trans. Amer. Math. Soc. 361 (2009), 2719-2754.

[16] Y. Lu and M.N. Do, *Multidimensional directional filterbanks and surfacelets* IEEE Trans. Image Process. **16** (2007) 918–931.

Compressive-wavefield simulations

Felix J. Herrmann, Yogi Erlangga, and Tim. T. Y. Lin

Department of Earth and Ocean Sciences, the University of British Columbia, Canada
fherrmann,yerlangga,tlin@eos.ubc.ca

Abstract:

Full-waveform inversion's high demand on computational resources forms, along with the non-uniqueness problem, the major impediment withstanding its widespread use on industrial-size datasets. Turning modeling and inversion into a compressive sensing problem—where simulated data are recovered from a relatively small number of independent simultaneous sources—can effectively mitigate this high-cost impediment. The key is in showing that we can design a sub-sampling operator that commutes with the time-harmonic Helmholtz system. As in compressive sensing, this leads to a reduction in simulation cost. Moreover, this reduction is commensurate with the transform-domain sparsity of the solution, implying that computational costs are no longer determined by the size of the discretization but by transform-domain sparsity of the solution of the CS problem which forms our data. The combination of this sub-sampling strategy with our recent work on implicit solvers for the Helmholtz equation provides a viable alternative to full-waveform inversion schemes based on explicit finite-difference methods.

1. Introduction

With the recent resurgence of full-waveform inversion—i.e., adjoint-state methods applied to solve PDE-constrained optimization problems—the computational cost of solving forward modeling has become one of the major impediments withstanding successful application of this technology to industry-size data volumes. To overcome this impediment, we argue that further improvements will depend on a problem formulation with a computational complexity that is no longer strictly determined by the *size* of the discretization but by transform-domain *sparsity* of its solution. In this new paradigm, we bring computational costs in par with our ability to compress solutions to certain PDEs. This premise is related to two recent developments. First, there is the new field of compressive sensing [CS in short throughout the paper, 4, 5]—where the argument is made, and rigorously proven—that compressible signals can be recovered from severely sub-Nyquist sampling by solving a sparsity promoting program. Second, there is in the seismic community the recent resurgence of simultaneous-source acquisition [1, 13, 2, 18, 12], and continuing efforts to reduce the cost of seismic modeling, imaging, and inversion through phase encoding of simultaneous sources [16, 21, 13, 12] and the removal of subsets of angular frequencies [22, 17, 15, 12] or plane waves [24]. All these approaches correspond to instances of CS. By using CS principles, we have been able to remove the associated sub-sampling interferences through a combination of exploiting transform-domain sparsity, properties of certain sub-sampling schemes, and the existence of sparsity promoting solvers.

2. Compressive full-waveform inversion

Full-waveform inversion entails solving PDE-constrained optimization problems of the following type:

$$\min_{\mathbf{U},\,\mathbf{m}} \frac{1}{2}\|\mathbf{R}\mathbf{M}\big(\mathbf{d}-\mathbf{D}\mathbf{U}\big)\|_2^2 \quad \text{s.t.} \quad \mathbf{H}[\mathbf{m}]\mathbf{U}=\mathbf{B}, \qquad (1)$$

where $\mathbf{d}$ and $\mathbf{U}$ are the observed data volumes and the solution of the multi-source (in its columns)-frequency Helmholtz equation over the domain of interest, $\mathbf{D}$ represents the detection operator that extracts the simulated data from time-harmonic solutions at the receiver locations, $\mathbf{H}$ a matrix with the discretized multi-frequency Helmholtz equation, and $\mathbf{B}$ a matrix with the frequency-transformed source distributions in its columns. In the above optimization problem (from which—after casting Eq. 1 in its unconstrained form—most quasi-Newton type full-waveform inversion schemes derive), solutions for the unknown velocity model, $\mathbf{m}$, and for the wave equation, $\mathbf{U}$, that minimize the energy mismatch are pursued. Because Eq. 1 is nonlinear in the model variables collected in the vector $\mathbf{m}$, solutions of Eq. 1 require multiple solves of the (implicit) Helmholtz equation. Even after preconditioning (yielding a complexity for this solver of $\mathcal{O}(n^4)$ in 2-D [7, 6]), this may prove computationally prohibitive. We address this problem by using CS [20, 12] to reduce the size of the seismic data volume through $\mathbf{y}=\mathbf{R}\mathbf{M}\mathbf{d}$ where

$$\mathbf{R}\mathbf{M}=\overbrace{\begin{bmatrix}\mathbf{R}_1^{\Sigma}\otimes\mathbf{I}\otimes\mathbf{R}_1^{\Omega}\\ \vdots \\ \mathbf{R}_{n_{s'}}^{\Sigma}\otimes\mathbf{I}\otimes\mathbf{R}_{n_{s'}}^{\Omega}\end{bmatrix}}^{\text{sub sampler}}\overbrace{\left(\mathbf{F}_2^{*}\,\mathrm{diag}\left(e^{\hat{\imath}\boldsymbol{\theta}}\right)\otimes\mathbf{I}\right)\mathbf{F}_3}^{\text{random phase encoder}},$$

with $\mathbf{F}_{2,3}$ the 2,3-D Fourier transforms, and $\boldsymbol{\theta}=\mathrm{Uniform}([0,2\pi])$ a random phase rotation. The matrices $\mathbf{R}^{\Omega}$ and $\mathbf{R}^{\Sigma}$ represent CS-subsampling matrices (see Figure 1) acting along the rows (frequency coordinate) and columns (source coordinate) of the data volume, respectively. As shown by [12] application of this CS-sampling

matrix, $\mathbf{RM}$, to the data is equivalent to applying it to the source wavefields directly, which turns single-impulsive sources into a smaller set ($n'_s \ll n_s$ with n_s the number of separated single-impulsive sources) of time-harmonic simultaneous sources that are randomly phase encoded and that have for each source-experiment a different set of angular frequencies missing—i.e., there are $n'_f \ll n_f$ (with n_f the number of frequencies of fully sampled data) frequencies non-zero (see Figure 1). This implies that the sub-sampling operator commutes with the Helmholtz system and this allows us to recast Eq. 1 into the following reduced form (consisting of fewer frequencies and fewer right-hand sides):

$$\min_{\underline{\mathbf{U}},\,\mathbf{m}} \frac{1}{2}\|\mathbf{y} - \underline{\mathbf{DU}}\|_2^2 \quad \text{s.t.} \quad \underline{\mathbf{H}}[\mathbf{m}]\underline{\mathbf{U}} = \underline{\mathbf{B}}, \qquad (2)$$

where the underlined quantities are related to the reduced Helmholtz system.

3. The time-harmonic Helmholtz system

Since their inception, iterative implicit matrix-free solutions to the time-harmonic Helmholtz equation have been plagued by lack of numerical convergence for decreasing mesh sizes and increasing angular frequencies [19]. The inclusion of deflation—a way to handle small eigenvalues that lead to slow convergence [7, 6]—can successfully remove this impediment, bringing 2- and 3-D solvers for the time-harmonic Helmholtz into reach. For a given source (right-hand side $\mathbf{b}$) and angular frequency ω ($:= 2\pi f$, with f the temporal frequency in Hz), the frequency-domain wavefield $\mathbf{u}$ is computed with a Krylov method that involves the following system of equations:

$$\boldsymbol{\mathcal{H}}[\omega]\boldsymbol{\mathcal{M}}^{-1}\boldsymbol{\mathcal{Q}}\hat{\mathbf{u}} = \mathbf{b}, \quad \mathbf{u} = \boldsymbol{\mathcal{M}}^{-1}\boldsymbol{\mathcal{Q}}\hat{\mathbf{u}},$$

where $\boldsymbol{\mathcal{H}}[\omega]$, $\boldsymbol{\mathcal{M}}$, and $\boldsymbol{\mathcal{Q}}$ represent the discretized monochromatic Helmholtz equation, the preconditioner, and the projection matrices, respectively. As shown by [8, 9], convergence is guaranteed by defining the preconditioning matrix $\boldsymbol{\mathcal{M}}$ in terms of the discretized shifted or damped Helmholtz operator $\mathcal{M} := -\nabla\cdot\nabla - \frac{\omega^2}{c(x)^2}(1 - \beta\hat{i})$, $\hat{i} = \sqrt{-1}$, with $\beta > 0$. With this preconditioning, the eigenvalues of $\boldsymbol{\mathcal{H}}\boldsymbol{\mathcal{M}}^{-1}$ are clustered into a circle in the complex plane. By the action of the projector matrix $\boldsymbol{\mathcal{Q}}$, these eigenvalues move towards unity on the real axis. These two operations lower the condition number, which explains the superior performance of this solver.

4. Source-solution CS-sampling equivalence

Aside from the required number of frequencies, the computational cost of full-wavefield simulation is determined by the number of sources—i.e., the number of right-hand sides. In the current simulation paradigm, the number of sources coincides with the number of single-impulsive source simulations. As prescribed by CS, this number can be reduced by designing a survey that consists of a relatively *small number* of simultaneous experiments with simultaneous sources that contain *subsets* of angular frequencies. Mathematically, we can accomplish this by applying a CS-sampling matrix, $\mathbf{RM}$, to the individual-impulsive sources collected in the vector $\mathbf{s}$. If we can show that the solution from this set of "compressed" sources $\underline{\mathbf{s}} = \mathbf{RMs}$, is identical to the compressively sampled solution yielded from modeling the *complete*, we are in the position to speed up our computations. This speed up is the result of a decreased number of experiments and angular frequencies that are present in the simultaneous source vector. For this to work, the solution $\underline{\mathbf{y}}$ must be equivalent to the solution $\mathbf{y}$, obtained by compressively sampling the full solution. More specifically, we need to demonstrate that the solutions for the full and compressed systems are equivalent—i.e., $\mathbf{y} = \underline{\mathbf{y}}$ in

$$\begin{cases} \mathbf{H} = \mathbf{D}^* \quad \underbrace{\mathbf{s}}_{\text{impulsive sources}} \\ \mathbf{HU} = \mathbf{B} \\ \mathbf{y} = \mathbf{RMDU} := \mathbf{RMd} \end{cases} \qquad \begin{cases} \underline{\mathbf{B}} = \underline{\mathbf{D}}^*\underline{\mathbf{s}} = \underline{\mathbf{D}}^* \quad \underbrace{\mathbf{RMs}}_{\text{sim. sources}} \\ \underline{\mathbf{HU}} = \underline{\mathbf{B}} \\ \underline{\mathbf{y}} = \underline{\mathbf{DU}}. \end{cases}$$

Here, $\mathbf{H} = \mathrm{diag}(\boldsymbol{\mathcal{H}}[\omega_i])$ is the block-diagonal discretized Helmholtz equation for each $\omega_i := 2\pi i \cdot \Delta f$, $i = 1 \cdots n_f$, with n_f the number of frequencies and Δf its sample interval. The adjoint (denoted by *) of the detection matrix $\mathbf{D}$ injects the individual sources into the multiple right-hand sides, $\mathbf{B} = [\mathbf{b}_1\ \mathbf{b}_2\ \cdots\ \mathbf{b}_{n_s}]$, with n_s the number of shots. This detection matrix extracts data at the receiver positions. Its adjoint inserts data at the co-located source positions. Each column of $\mathbf{U}$ contains the wavefields for all frequencies induced by the shots located in the columns of $\mathbf{B}$. Consequently, the full simulation requires the inversion of the block-diagonal system (for all shots), followed by a detection—i.e., we have $\mathbf{d} = \mathbf{DH}^{-1}\mathbf{B}$, with $\mathbf{H}^{-1} = \mathrm{diag}(\boldsymbol{\mathcal{H}}^{-1}[\omega_i])$, $i = 1 \cdots n_s$. After CS sampling, this volume is reduced to $\mathbf{y} = \mathbf{RMd}$ by applying the flat rectangular CS-sampling matrix $\mathbf{RM}$ (defined explicitly in the next section) to the full simulation. Applying $\mathbf{RM}$ directly to the sources $\mathbf{s}$ leads to a compressed system $\underline{\mathbf{H}}$, which after inversion gives $\underline{\mathbf{y}}$. To illustrate why $\mathbf{y}$ is equivalent to $\underline{\mathbf{y}}$, consider a compressive sampling of the solution over frequency by the subsampling matrix $\mathbf{R}^\Omega$ (for clarity, we removed the orthonormal measurement matrix). This restriction matrix removes arbitrary rows from the right-hand side. By virtue of the block-diagonal structure of our system, we have $\mathbf{R}^\Omega\mathbf{H}^{-1} = \underline{\mathbf{H}}^{-1}\mathbf{R}^\Omega$ with $\underline{\mathbf{H}}^{-1} = \mathrm{diag}(\boldsymbol{\mathcal{H}}^{-1}[\omega_i]), i \subset \{1 \cdots n_f\}$, yielding $\mathbf{R}^\Omega\mathbf{U} = \underline{\mathbf{H}}^{-1}\underline{\mathbf{B}} = \underline{\mathbf{U}}$, where $\underline{\mathbf{B}} := \mathbf{R}^\Omega\mathbf{B}$. This means that frequency subsampling the right-hand side, followed by solving the system for the corresponding frequencies, is the same as solving the full system, followed by frequency subsampling. A similar argument holds when subsampling the shots (removing arbitrary columns of $\mathbf{B}$). Now, we have the reduced system $\mathbf{R}^\Omega\mathbf{U}(\mathbf{R}^\Sigma)^* = \underline{\mathbf{H}}^{-1}\underline{\mathbf{B}} = \underline{\mathbf{U}}$, with $\underline{\mathbf{B}} := \mathbf{R}^\Omega\mathbf{B}(\mathbf{R}^\Sigma)^*$. Using Kronecker products, these relations can be written succinctly as $(\mathbf{R}^\Sigma \otimes \mathbf{R}^\Omega)\mathrm{vec}\,(\mathbf{U}) = \mathrm{vec}\,(\underline{\mathbf{U}})$ and $(\mathbf{R}^\Sigma \otimes \mathbf{R}^\Omega)\mathrm{vec}\,(\mathbf{B}) = \mathrm{vec}\,(\underline{\mathbf{B}})$ with $\mathrm{vec}\,(\cdot)$ being a linear operator that maps a matrix into a lexicographically-sorted array. The inversion of $\underline{\mathbf{HU}} = \underline{\mathbf{B}}$ is easier because it involves only a subset of angular frequencies and simultaneous shots—i.e., $\{\underline{\mathbf{U}}, \underline{\mathbf{B}}\}$ contain only n'_s columns with n'_f frequency components each. Finally, the matrix $\underline{\mathbf{D}}$ extracts the compressed data from the solution.

5. Recovery by sparsity promotion

Aside from CS sampling the recovery from simultaneous simulations depends on a sparsifying transform that compresses seismic data, is fast, and reasonably incoherent with the CS sampling matrix. We accomplish this by defining the sparsity transform as the Kronecker product between the 2-D discrete curvelet transform [3] along the source-receiver coordinates, and the discrete wavelet transform along the time coordinate—i.e., $\mathbf{S} := \mathbf{C} \otimes \mathbf{W}$ with $\mathbf{C}$, $\mathbf{W}$ the curvelet- and wavelet-transform matrices, respectively. We reconstruct the seismic wavefield by solving the following nonlinear optimization problem

$$\widetilde{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{y}, \qquad (3)$$

with $\widetilde{\mathbf{d}} = \mathbf{S}^*\widetilde{\mathbf{x}}$ the reconstruction, $\mathbf{A} := \mathbf{R}\mathbf{M}\mathbf{S}^*$ the CS matrix, and $\mathbf{y}\,(=\underline{\mathbf{y}})$ the compressively simulated data (cf. Equation 2-right). Equation 3 is solved by SPGℓ_1 [23], a projected-gradient algorithm with root finding.

6. Computational complexity analysis

According to [19], the cost of the iterative Helmholtz solver equals $n_f n_s n_{it} \mathcal{O}(n^d)$, typically with $n_{it} = \mathcal{O}(n)$ the number of iterations. For $d = 2$ and assuming $n_s = n_f = \mathcal{O}(n)$, this cost becomes $\mathcal{O}(n^5)$. Under the same assumption, the cost of a time-domain solver is $\mathcal{O}(n^4)$. The iterative Helmholtz solver can only become competitive if $n_{it} = \mathcal{O}(1)$, yielding an $\mathcal{O}(n^4)$ computational complexity. [7, 6] achieve this by the method explained earlier. Despite this improvement, this figure is still overly pessimistic for simulations that permit sparse representations. As long as the simulation cost exceeds the ℓ_1-recovery cost (cf. Equation 3), CS will improve on this result. This reduction depends on the cost of $\mathbf{A}$, which is dominated by the CS-matrix. For naive choices, such as Gaussian projections, these sampling matrices cost $\mathcal{O}(n^3)$ for each frequency, which offers no gain. However, with our choice of fast $\mathcal{O}(n \log n)$ projections with random convolutions [20], we are able to reduce this cost to $\mathcal{O}(n^2 \log n)$. Note that these costs are of the same order as those of calculating the sparsifying transforms. Now, the leading order cost of the ℓ_1 recovery is reduced to $\mathcal{O}(n^3 \log n)$, which is significantly less than the cost of solving the full Helmholtz system, especially for large problems ($n \to \infty$) and for extensions to $d = 3$.

7. Example

To illustrate CS-recovery quality, we conduct a series of experiments for two velocity models, namely the complex model used in [10], and a simple single-layer model. These models generate seismic lines that differ in complexity. During these experiments, we vary the subsampling ratio and the frequency-to-shot subsampling ratio. All simulations are carried out with a fully parallel Helmholtz solver for a spread with 128 collocated shots and receivers sampled at a 30 m interval. The time sample interval is $0.004\,\mathrm{s}$ and the source function is a Ricker wavelet with a central frequency of 10 Hz. By solving Equation 3, we recover the full simulation for the two datasets. Comparison between the full and compressive simulations in Figure 2 shows remarkable high-fidelity results even for increasing subsampling ratios. As expected, the SNR for the simple model is better because of the reduced complexity, whereas the numbers in Table 1 for the complex model confirm increasing recovery errors for increasing subsampling ratios. Moreover, the bandwidth limitation of seismic data explains improved recovery with decreasing frequency-to-shot ratio for a fixed subsampling ratio. Because the speedup of the solution is roughly proportional to the subsampling ratio, we can conclude that speedups of four to six times are possible with a minor drop in SNR.

Subsample ratio	0.25	0.15	0.07
n'_f/n'_s	recovery error (dB)		
2	14.3	12.1	8.6
1	18.2	14.5	10.2
0.5	22.2	16.5	10.7
Speed up (%)	400	670	1420

Table 1: Signal-to-noise ratios based on the complex model, $\mathrm{SNR} = -20\log_{10}(\frac{\|\mathbf{d}-\widetilde{\mathbf{d}}\|_2}{\|\mathbf{d}\|_2})$ for reconstructions with the curvelet-wavelet sparsity transform for different subsample and frequency-to-shot ratios..

8. Discussion, extensions, and conclusions

Compressive sampling (CS) can be considered a paradigm shift because objects of interest that exhibit transform-domain sparsity can be recovered from degrees of subsampling commensurate their sparsity. This new paradigm can be applied to reduce the computational complexity of solving PDEs that lie at the heart of PDE-constrained optimization problems. In this paper, we demonstrate that this principle leads to simultaneous source experiments that reduce the cost of computer simulations. Similar cost reductions are possible during actual acquisition in situations where we have control over the physical sources; such as during acquisition on land [14]. These results are exciting because CS decouples simulation- and acquisition-related costs from the discretization size. Instead, these costs depend on sparsity. Because the image space is even sparser after focusing seismic energy, we obtain further improvements when we extend CS principles to promote joint sparsity through mixed $(1,\,2)$-norm minimization [11].

References

[1] Craig J. Beasley. A new look at marine simultaneous sources. *The Leading Edge*, 27(7):914–917, 2008.

[2] A. J. Berkhout. Changing the mindset in seismic data acquisition. *The Leading Edge*, 27(7):924–938, 2008.

[3] E. J. Candès, L. Demanet, D. L. Donoho, and L. Ying. Fast discrete curvelet transforms. *Multiscale Modeling and Simulation*, 5:861–899, 2006.

[4] E.J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

[5] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[6] Y. A. Erlangga and F. J. Herrmann. An iterative multilevel method for computing wavefields in frequency-domain seismic inversion. In *SEG Technical Program Expanded Abstracts*, volume 27, pages 1957–1960. SEG, November 2008.

[7] Y A Erlangga and R Nabben. On multilevel projection Krylov method for the preconditioned Helmholtz system. 2007. Submitted for publication.

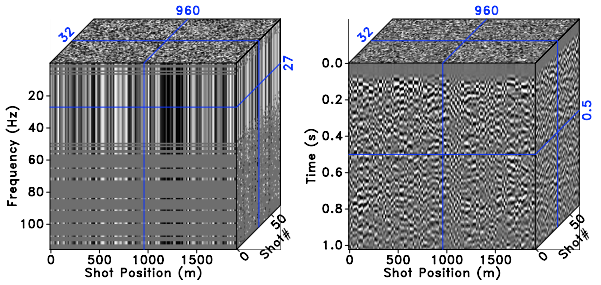


Figure 1: Compressive sampling with simultaneous sources. **(a)** Amplitude spectrum for the source signatures emitted by each source as part of the simultaneous-source experiments. These signatures appear noisy in the shot-receiver coordinates because of the phase encoding (cf. Equation 1). Observe that the frequency restrictions are different for each simultaneous source experiment. **(b)** CS-data after applying the inverse Fourier transform. Notice the noisy character of the simultaneous-shot interferences..

[8] Y A Erlangga, C Vuik, and C W Oosterlee. On a class of preconditioners for solving the Helmholtz equation. *Applied Numerical Mathematics*, 50:409–425, 2004.

[9] Y A Erlangga, C Vuik, and C W Oosterlee. Comparison of multigrid and incomplete LU shifted-Laplace preconditioners for the inhomogeneous Helmholtz equation. *Applied Numerical Mathematics*, 56:648–666, 2006.

[10] F. J. Herrmann, U. Boeniger, and D. J. Verschuur. Non-linear primary-multiple separation with directional curvelet frames. *Geophysical Journal International*, 170:781–799, 2007.

[11] Felix J. Herrmann. Compressive imaging by wavefield inversion with group sparsity. Technical Report TR-2009-01, UBC-SLIM, 2009.

[12] Felix J. Herrmann, Yogi A. Erlangga, and Tim T.Y. Lin. Compressive simultaneous full-waveform simulation. TR-2008-09. to appear in geophysics. 2009.

[13] C.E. Krohn and R. Neelamani. Simultaneous sourcing without compromise. In *Rome 2008, 70th EAGE Conference & Exhibition*, page B008, 2008.

[14] Tim T Y Lin and Felix J Herrmann. Designing simultaneous acquisitions with compressive sensing. In *Amsterdam 2009, 71th EAGE Conference & Exhibition*, 2009.

[15] T.T.Y. Lin, E. Lebed, Y. A. Erlangga, and F. J. Herrmann. Interpolating solutions of the helmholtz equation with compressed sensing. In *SEG Technical Program Expanded Abstracts*, volume 27, pages 2122–2126. SEG, November 2008.

[16] S. A. Morton and C. C. Ober. Faster shot-record depth migrations using phase encoding. In *SEG Technical Program Expanded Abstracts*, volume 17, pages 1131–1134. SEG, 1998.

[17] W. Mulder and R. Plessix. How to choose a subset of frequencies in frequency-domain finite-difference migration. 158:801–812, 2004.

[18] N. Neelamani, C. Krohn, J. Krebs, M. Deffenbaugh, and J. Romberg. Efficient seismic forward modeling using simultaneous random sources and sparsity. In *SEG International Exposition and 78th Annual Meeting*, pages 2107–2110, 2008.

[19] C. D. Riyanti, Y. A. Erlangga, R.-E. Plessix, W. A. Mulder, C. Vuik, and C. Oosterlee. A new iterative solver for the time-harmonic wave equation. *Geophysics*, 71(5):E57–E63, 2006.

[20] J. Romberg. Compressive sensing by random convolution. *submitted*, 2008.

[21] L. A. Romero, D. C. Ghiglia, C. C. Ober, and S. A. Morton. Phase encoding of shot records in prestack migration. *Geophysics*, 65(2):426–436, 2000.

[22] Laurent Sirgue and R. Gerhard Pratt. Efficient waveform inversion and imaging: A strategy for selecting temporal frequencies. *Geophysics*, 69(1):231–248, 2004.

[23] E. van den Berg and M. P. Friedlander. Probing the pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, 31(2):890–912, 2008.

[24] Denes Vigh and E. William Starr. 3d prestack plane-wave, full-waveform inversion. *Geophysics*, 73(5):VE135–VE144, 2008.

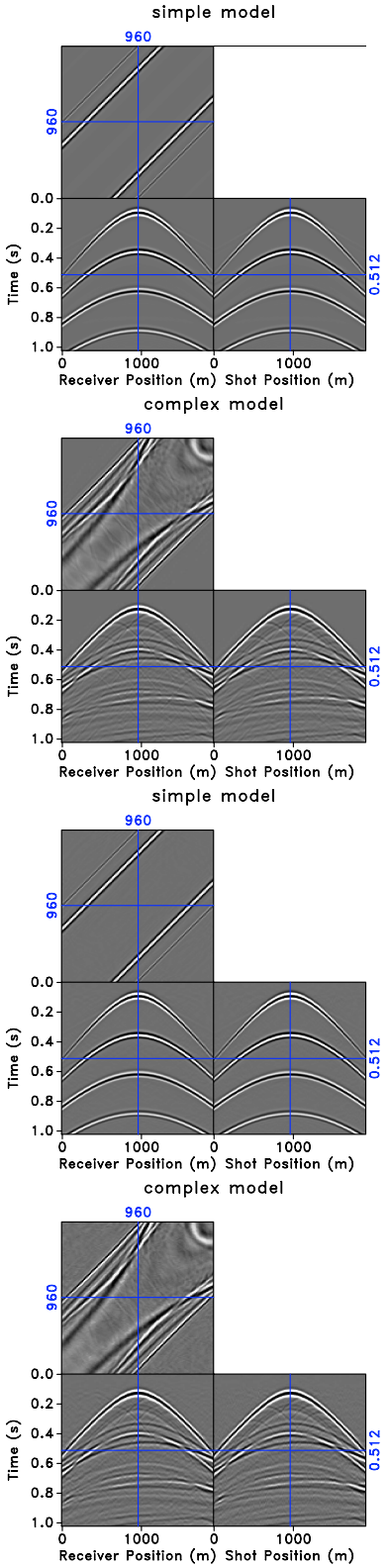


Figure 2: Comparison between conventional and compressive simulations in for simple and complex velocity models. **(a)** Crossing-planes view of the seismic line for the simple model. **(b)** The same for the complex model. **(c)**. Recovered simulation (with a SNR of 28.1 dB) for the simple model from 25 % of the samples with the ℓ_1-solver running to convergence. **(d)** The same but for the complex model now with a SNR of 18.2 dB..

Computable Fourier Conditions for Alias-Free Sampling and Critical Sampling

Yue M. Lu [(1)(2)], Minh N. Do [(2)] and Richard S. Laugesen [(2)]

(1) Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015, Lausanne, Switzerland
(2) University of Illinois at Urbana-Champaign, Urbana IL 61801, USA
yue.lu@epfl.ch, minhdo@illinois.edu, Laugesen@illinois.edu

Abstract:

We propose a Fourier analytical approach to the problems of alias-free sampling and critical sampling. Central to this approach are two Fourier conditions linking the above sampling criteria with the Fourier transform of the indicator function defined on the underlying frequency support. We present several examples to demonstrate the usefulness of the proposed Fourier conditions in the design of critically sampled multidimensional filter banks. In particular, we show that it is impossible to implement any cone-shaped frequency partitioning by a nonredundant filter bank, except for the 2-D case.

1 Introduction

The search for *alias-free sampling* lattices for a given frequency support, and in particular for those lattices achieving minimum sampling densities, is a fundamental issue in various signal processing applications that involve the design of efficient acquisition schemes for bandlimited signals. As a special case of alias-free sampling, the concept of *critical sampling* also plays an important role in the theory and design of critically sampled (a.k.a. maximally decimated) multidimensional filter banks [9].

The study of alias-free (and critical) sampling lattices is a classical problem [8, 4]. So far, most existing work in the literature approaches the problem from a geometrical perspective: The primary tools employed include the theories from Minkowski's work [2], as well as various geometrical intuitions and heuristics.

In this paper, we propose a Fourier analytical approach to the problems of alias-free sampling and critical sampling. Central to this approach are two Fourier conditions linking the above sampling criteria with the Fourier transform of the indicator function defined on the underlying frequency support (see Theorem 1 and Proposition 2). An important feature of the proposed conditions is that they open the door to purely analytical and computational solutions to the sampling lattice selection problem.

The rest of the paper is organized as follows. In Section 2, we briefly review some relevant concepts on sampling bandlimited signals. We present in Section 3 a novel condition linking the alias-free sampling (as well as critical sampling) with the Fourier transform of the indicator function defined on the given frequency support. In Section 4, we present an application of the proposed Fourier conditions in the design of multidimensional nonredundant filter banks. We conclude the paper in Section 5. The material in this paper was presented in part in [5] and [7]. As a novel aspect, we present in this paper a different proof for Theorem 1, which provides important new insights into this key result.

Notation: The Fourier transform of a function $f(\boldsymbol{\omega})$ defined on $\mathbb{R}^N$ is defined by

$$\widehat{f}(\boldsymbol{x}) = \int_{\mathbb{R}^N} f(\boldsymbol{\omega})\, e^{-2\pi j\, \boldsymbol{x}\cdot\boldsymbol{\omega}}\, d\boldsymbol{\omega}. \tag{1}$$

Calligraphic letters, such as $\mathcal{D}$, represent bounded and open frequency domains in $\mathbb{R}^N$, with $m(\mathcal{D})$ denoting the Lebesgue measure (*i.e.* volume) of $\mathcal{D}$. Given a nonsingular matrix $\boldsymbol{M}$ and a vector $\boldsymbol{\tau}$, we use $\boldsymbol{M}(\mathcal{D}+\boldsymbol{\tau})$ to represent the set of points of the form $\boldsymbol{M}(\boldsymbol{\omega}+\boldsymbol{\tau})$ for $\boldsymbol{\omega}\in\mathcal{D}$. Finally, we denote by $\mathbb{1}_{\mathcal{D}}(\boldsymbol{\omega})$ the indicator function of the domain $\mathcal{D}$, *i.e.*, $\mathbb{1}_{\mathcal{D}}(\boldsymbol{\omega})=1$ if $\boldsymbol{\omega}\in\mathcal{D}$ and $\mathbb{1}_{\mathcal{D}}(\boldsymbol{\omega})=0$ otherwise.

2 Background

In multidimensional multirate signal processing, the sampling operations are usually defined on lattices, each of which can be generated by an $N\times N$ nonsingular matrix $\boldsymbol{M}$ as

$$\Lambda_{\boldsymbol{M}} \stackrel{\text{def}}{=} \{\boldsymbol{M}\boldsymbol{n} : \boldsymbol{n}\in\mathbb{Z}^N\}. \tag{2}$$

We denote by $\Lambda^*_{\boldsymbol{M}}$ the corresponding reciprocal lattice (a.k.a. polar lattice), defined as

$$\Lambda^*_{\boldsymbol{M}} \stackrel{\text{def}}{=} \{\boldsymbol{M}^{-T}\boldsymbol{\ell} : \boldsymbol{\ell}\in\mathbb{Z}^N\} \tag{3}$$

In the rest of the paper, when it is clear from the context what the generating matrix is, we will drop the subscripts in $\Lambda_{\boldsymbol{M}}$ and $\Lambda^*_{\boldsymbol{M}}$, and use Λ and Λ^* for simplicity.

Let $f(\boldsymbol{x})$ be a continuous-domain signal, whose Fourier transform is bandlimited to a bounded open set $\mathcal{D}\subset\mathbb{R}^N$. The discrete-time Fourier transform of the samples $s[\boldsymbol{n}]\stackrel{\text{def}}{=} f(\boldsymbol{M}\boldsymbol{n})$ is supported in [9]

$$\mathcal{S} = \boldsymbol{M}^T\left(\bigcup_{\boldsymbol{k}\in\Lambda^*}(\mathcal{D}+\boldsymbol{k})\right). \tag{4}$$

For appropriately chosen sampling lattices, the aliasing components in (4) do not overlap with the baseband frequency

support $\mathcal{D}$. In this important case, we can fully recover the original continuous-domain signal $f(\boldsymbol{x})$ by applying an ideal interpolation filter spectrally supported on $\mathcal{D}$ to the discrete samples $s[\boldsymbol{n}]$.

Definition 1 *We say a frequency support $\mathcal{D}$ allows an* alias-free *$\boldsymbol{M}$-fold sampling, if different shifted copies of $\mathcal{D}$ in* (4) *are disjoint, i.e.,*

$$\mathcal{D} \cap (\mathcal{D} + \boldsymbol{k}) = \emptyset \;\; \textit{for all} \;\; \boldsymbol{k} \in \Lambda^* \setminus \{\boldsymbol{0}\}. \tag{5}$$

Furthermore, we say $\mathcal{D}$ can be critically sampled *by $\boldsymbol{M}$, if in addition to the alias-free condition in* (5)*, the union of the shifted copies also covers the entire spectrum, i.e.,*

$$\bigcup_{\boldsymbol{k} \in \Lambda^*} (\mathcal{D} + \boldsymbol{k}) = \mathbb{R}^N, \quad \textit{up to a set of measure zero}. \tag{6}$$

The focus of this work is to present two Fourier analytical conditions for alias-free sampling and critical sampling. Our discussions will be based on the following geometrical argument [2], which can be easily verified from (5).

Proposition 1 *The alias-free sampling condition in* (5) *is equivalent to requiring*

$$\Lambda^* \cap (\mathcal{D} - \mathcal{D}) = \{\boldsymbol{0}\}, \tag{7}$$

where $\mathcal{D} - \mathcal{D} \overset{def}{=} \{\boldsymbol{\omega} - \boldsymbol{\tau} : \boldsymbol{\omega}, \boldsymbol{\tau} \in \mathcal{D}\}$ is the Minkowski sum of the open set $\mathcal{D}$ and its negative $-\mathcal{D}$.

3 Fourier Analytical Conditions

In this section, we study the problems of alias-free sampling and critical sampling with Fourier techniques. The key observation is a link between the alias-free sampling condition and the Fourier transform of the indicator function $\mathbb{1}_{\mathcal{D}}(\boldsymbol{\omega})$ defined on the frequency support $\mathcal{D}$.

31 Alias-Free Sampling

Lemma 1 *Let $\mathcal{D}$ be a frequency region, and $f(\boldsymbol{\omega})$ a positive function supported on $(\mathcal{D} - \mathcal{D})$, i.e., $f(\boldsymbol{\omega}) > 0$ for $\boldsymbol{\omega} \in (\mathcal{D} - \mathcal{D})$ and $f(\boldsymbol{\omega}) = 0$ otherwise. Then $\mathcal{D}$ allows an $\boldsymbol{M}$-fold alias-free sampling* if and only if

$$\sum_{\boldsymbol{k} \in \Lambda^*} f(\boldsymbol{k}) = f(\boldsymbol{0}). \tag{8}$$

Proof By construction, (8) holds if and only if $\Lambda^* \cap (\mathcal{D} - \mathcal{D}) = \{\boldsymbol{0}\}$. Applying Proposition 1, we are done. ∎

Theorem 1 *A frequency region $\mathcal{D}$ allows an $\boldsymbol{M}$-fold alias-free sampling* if and only if

$$|\boldsymbol{M}| \sum_{\boldsymbol{n} \in \Lambda} |\widehat{\mathbb{1}}_{\mathcal{D}}(\boldsymbol{n})|^2 = m(\mathcal{D}), \tag{9}$$

where $\widehat{\mathbb{1}}_{\mathcal{D}}(\boldsymbol{x})$ is the Fourier transform of $\mathbb{1}_{\mathcal{D}}(\boldsymbol{\omega})$, and $|\boldsymbol{M}|$ is the absolute value of the determinant of $\boldsymbol{M}$.

Proof Consider the autocorrelation function

$$R_{\mathcal{D}}(\boldsymbol{\omega}) = \int \mathbb{1}_{\mathcal{D}}(\boldsymbol{\tau})\, \mathbb{1}_{\mathcal{D}}(\boldsymbol{\tau} - \boldsymbol{\omega})\, d\boldsymbol{\tau}.$$

Clearly, $R_{\mathcal{D}}(\boldsymbol{\omega}) \geq 0$ for all $\boldsymbol{\omega}$. Meanwhile, we can verify that supp $R_{\mathcal{D}}(\boldsymbol{\omega}) = (\mathcal{D} - \mathcal{D})$. Thus, we can apply Lemma 1 and obtain that, $\mathcal{D}$ allows an $\boldsymbol{M}$-fold alias-free sampling if and only if

$$\sum_{\boldsymbol{k} \in \Lambda^*} R_{\mathcal{D}}(\boldsymbol{k}) = R_{\mathcal{D}}(\boldsymbol{0}) = \int \mathbb{1}_{\mathcal{D}}(\boldsymbol{\tau})\, d\boldsymbol{\tau} = m(\mathcal{D}).$$

Applying the Poisson summation formula to the above equality (see Appendix A of [7] for a justification of the pointwise equality), we have

$$m(\mathcal{D}) = \sum_{\boldsymbol{k} \in \Lambda^*} R_{\mathcal{D}}(\boldsymbol{k}) = |\boldsymbol{M}| \sum_{\boldsymbol{n} \in \Lambda} \widehat{R}_{\mathcal{D}}(\boldsymbol{n}). \tag{10}$$

From the definition of $R_{\mathcal{D}}(\boldsymbol{\omega})$, its Fourier transform is $\widehat{R}_{\mathcal{D}}(\boldsymbol{x}) = |\widehat{\mathbb{1}}_{\mathcal{D}}(\boldsymbol{x})|^2$. Substituting this formula into (10), we are done. ∎

32 Critical Sampling

Here we focus on the special case of critical sampling, and begin by mentioning, without proof, a standard result:

Lemma 2 *A frequency support $\mathcal{D}$ can be critically sampled by a sampling matrix $\boldsymbol{M}$ if and only if $\boldsymbol{M}$ is an alias-free sampling matrix for $\mathcal{D}$ with sampling density $1/|\boldsymbol{M}| = m(\mathcal{D})$.*

Proposition 2 *A frequency support $\mathcal{D}$ can be critically sampled by a matrix $\boldsymbol{M}$ if and only if*

$$\widehat{\mathbb{1}}_{\mathcal{D}}(\boldsymbol{0}) = m(\mathcal{D}) = \frac{1}{|\boldsymbol{M}|} \quad \textit{and} \quad \widehat{\mathbb{1}}_{\mathcal{D}}(\boldsymbol{n}) = 0 \tag{11}$$

for all $\boldsymbol{n} \in \Lambda \setminus \{\boldsymbol{0}\}$.

Proof Suppose (11) holds. Then it follows that

$$\sum_{\boldsymbol{n} \in \Lambda} |\widehat{\mathbb{1}}_{\mathcal{D}}(\boldsymbol{n})|^2 = |\widehat{\mathbb{1}}_{\mathcal{D}}(\boldsymbol{0})|^2 = \frac{m(\mathcal{D})}{|\boldsymbol{M}|},$$

and hence from Theorem 1, $\boldsymbol{M}$ is an alias-free sampling matrix for $\mathcal{D}$. Meanwhile, since $m(\mathcal{D}) = \frac{1}{|\boldsymbol{M}|}$, we can apply Lemma 2 to conclude that $\mathcal{D}$ is critically sampled by $\boldsymbol{M}$. By reversing the above line of reasoning, we can also show the necessity of (11). ∎

Remark: The result of Proposition 2 is previously known in various disciplines. In approximation theory, the condition (11) is often called the interpolation property (see, for example, [4]). The usefulness of this condition in the context of lattice tiling was first pointed out by Kolountzakis and Lagarias [3] and applied to investigate the tiling of various high dimensional shapes.

33 Computational Aspects

The Fourier conditions proposed in Theorem 1 and Proposition 2 can lead to practical computational algorithms for testing alias-free and critical sampling. Here, we briefly comment on two important computational aspects in applying the proposed conditions.

First, as a prerequisite to using the proposed Fourier conditions, we must know the expression for $\widehat{\mathbb{1}}_{\mathcal{D}}(\boldsymbol{x})$. This evaluation can be a cumbersome task if we need to do the derivation by hand for each given $\mathcal{D}$. However, when the frequency regions $\mathcal{D}$ are arbitrary polygonal and polyhedral domains, we can obtain the closed-form expressions for $\widehat{\mathbb{1}}_{\mathcal{D}}(\boldsymbol{x})$ via the divergence theorem [1, 7].

Another potential issue in practical implementations is that the Fourier conditions in (9) and (11) both involve an infinite number of lattice points. We show in [7] that the infinite sum in (9) can be well-approximated by a truncated finite sum. Moreover, with high probability, we actually only need to evaluate the Fourier transform on a very small number of points in a lattice (e.g. 4 points in 2-D) in order to show aliasing occurs, thus ruling out the lattice.

4 Application: Filter Bank Design

In this section we present an application of Proposition 2 in the design of multidimensional critically sampled filter banks.

41 Frequency Partitioning of Critically Sampled Filter Banks

Consider a general multidimensional filter bank, where each channel contains a subband filter and a sampling operator. As an important step in filter bank design, we need to specify the ideal passband support of each subband filter, all of which form a partitioning of the frequency spectrum.

Not every possible frequency partitioning can be used for filter bank implementation though. In particular, if we want to have a nonredundant filter bank, then the ideal passband support of each subband filter must be critically sampled by the sampling matrix in that channel. Consequently, whenever given a possible frequency partitioning, we must first perform a "reality check" of seeing whether the above condition is met, before proceeding to actual filter design.

The critical sampling condition is commonly verified geometrically (*i.e.* by drawing figures). Although intuitive and straightforward, this geometrical approach becomes cumbersome when the shape of the passband support is complicated, or when we work in 3-D and higher dimensional cases. Applying the result of Proposition 2, we propose in the following a computational procedure, which can systematically check and determine the critical sampling matrices of a given polytope region. Notice that the algorithm only searches among integer matrices, since the filter banks considered here operate on discrete-time signals.

Procedure 1 *Let $\mathcal{D}$ be a given polytope-shaped frequency support region.*

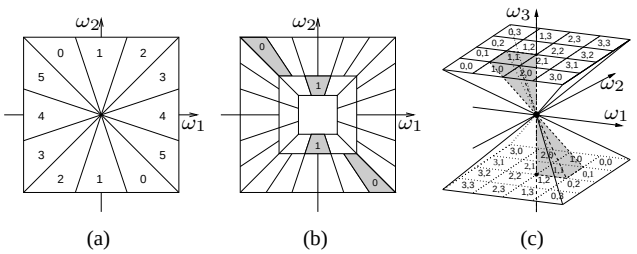


Figure 1: The ideal frequency partitioning of several filter banks. (a) A directional filter bank which decomposes the frequency cell $(-\frac{1}{2}, \frac{1}{2}]^2$ into 6 subbands. (b) A directional multiresolution frequency partitioning. (c) A 3-D directional frequency decomposition with pyramid-shaped passband supports.

1. *Calculate $\delta = 1/m(\mathcal{D})$. From* (11)*, any matrix $\boldsymbol{M}$ that can critically-sample $\mathcal{D}$ must satisfy $|\boldsymbol{M}| = \delta$. If δ is not an integer, then stop the procedure, since in this case it is impossible for $\mathcal{D}$ to be critically sampled by any integer matrix.*
2. *Construct a closed-form formula [7] for $\widehat{\mathbb{1}}_{\mathcal{D}}(\boldsymbol{x})$.*
3. *Based on the Hermite normal form, construct an exhaustive list of matrices of determinant δ, each corresponding to a distinct sampling lattice [7].*
4. *For every matrix $\boldsymbol{M}$ in the above list, test the following condition*

$$\widehat{\mathbb{1}}_{\mathcal{D}}(\boldsymbol{M}\boldsymbol{n}) = 0 \quad \text{for all } \boldsymbol{n} \in \mathbb{Z}^N \setminus \{\boldsymbol{0}\} \text{ with } \|\boldsymbol{n}\|_\infty \le r, \tag{12}$$

   where r is a large positive integer.
5. *Present all the matrices in the list that satisfy* (12)*. If there is no such matrix, then $\mathcal{D}$ cannot be critically sampled by any integer matrix.*

To be clear, the expression (12) is a necessary condition for $\mathcal{D}$ to be critically sampled by $\boldsymbol{M}$. It is not sufficient since we only check for integer points within a finite radius r, and so in principle, even if $\boldsymbol{M}$ satisfies (12) for all $\|\boldsymbol{n}\|_\infty \le r$, it might happen that $\widehat{\mathbb{1}}_{\mathcal{D}}(\boldsymbol{M}\boldsymbol{n}) \neq 0$ for some $\boldsymbol{n}$ with $\|\boldsymbol{n}\|_\infty > r$. However, by choosing r sufficiently large, we can gain confidence in the validity of the original infinite condition (11) as required in Proposition 2. We leave the quantitative analysis of this approximation to [7]. In the following examples, we choose $r = 10000$.

Example 1 *Figure 1(a) presents the frequency decomposition of a directional filter bank (DFB). Applying the algorithm in Procedure 1, we can easily verify that this frequency decomposition can be critically sampled. The corresponding sampling matrices, denoted by $\boldsymbol{M}_k$ for the kth subband, are*

$$\boldsymbol{M}_0 = \boldsymbol{M}_1 = \boldsymbol{M}_2 = \begin{pmatrix} 6 & 3 \\ 0 & 1 \end{pmatrix}.$$

$\boldsymbol{M}_3, \boldsymbol{M}_4$ and $\boldsymbol{M}_5$ can be inferred by symmetry.

Example 2 *We show in Figure 1(b) a directional and multiresolution decomposition of the 2-D frequency spectrum. Applying Procedure 1 confirms that such a frequency partitioning can be critically sampled as well. The sampling*

matrices for two representative subbands (marked as dark regions in the figure) are

$$\boldsymbol{M}_0 = \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix} \text{ and } \boldsymbol{M}_1 = \begin{pmatrix} 8 & 4 \\ 0 & 4 \end{pmatrix}.$$

Example 3 *Figure 1(c) shows an extension of the original 2-D DFB to the 3-D case [6]. Applying Procedure 1, we find that the 3-D frequency partitioning shown in Figure 1(c) cannot be critically sampled; in other words, redundancy is unavoidable for a 3-D DFB.*

42 Critical Sampling of General Cone-Shaped Frequency Regions in Higher Dimensions

The result in Example 3 can be generalized to higher dimensions, and to cases where the subbands take different directional shapes. As an application of the Fourier condition in Proposition 2, we show here a much more general statement: it is impossible to implement any cone-shaped frequency partitioning by a nonredundant filter bank, except for the 2-D case.

We consider the following ideal subband supports in N-D:

$$\mathcal{D} = \{\boldsymbol{\omega} : a \le |\omega_N| \le b,\ (\omega_1, \ldots, \omega_{N-1}) \in \omega_N\, \mathcal{B}\}, \quad (13)$$

where $\mathcal{B}$ is some bounded set in $\mathbb{R}^{N-1}$. Geometrically, $\mathcal{D}$ takes the form of a two-sided cone in $\mathbb{R}^N$, truncated by hyperplanes $|\omega_N| = a$ and $|\omega_N| = b$, where $0 \le a < b$. The "base" region $\mathcal{B}$ in (13) is the intersection between the cone and the hyperplane $\omega_N = 1$.

The formulation in (13) is flexible enough to characterize, up to a rotation, any directional subband shown in Figure 1. For example, the 3-D pyramid-shaped subband $(1,1)$ in Figure 1(c) can be presented by $a = 0, b = \frac{1}{2}$, and $\mathcal{B} = [-\frac{1}{2}, 0]^2$. However, the class of frequency shapes that can be described by (13) is far beyond those shown in Figure 1, since the formulation (13) allows for arbitrary configuration of the cross section heights a and b (not necessarily the dyadic decomposition as in Figure 1(b)) and arbitrary shape for the base $\mathcal{B}$ (not necessarily lines or squares).

Lemma 3 *If a frequency support $\mathcal{D}$ can be critically sampled by an* integer *matrix $\boldsymbol{M}$, then*

$$\widehat{\mathbb{1}}_{\mathcal{D}}(|\boldsymbol{M}|\,\boldsymbol{n}) = 0, \quad \textit{for all } \boldsymbol{n} \in \mathbb{Z}^N \setminus \{0\}. \quad (14)$$

Proof It is easy to verify that, for any integer matrix $\boldsymbol{M}$, the vector $|\boldsymbol{M}|\,\boldsymbol{n}$ belongs to the lattice Λ generated by $\boldsymbol{M}$. The condition (14) then follows from (11) in Proposition 2. ∎

Theorem 2 *For arbitrary choice of $0 \le a < b$ and the base shape $\mathcal{B}$, the frequency domain support $\mathcal{D}$ given in* (13) *cannot be critically sampled by any integer matrix in N-dimensions, $N \ge 3$.*

Remark: For 2-D, we established the positive result in Examples 1 and 2.

Proof We argue by contradiction. Suppose for $N \ge 3$, and for some particular choices of $0 \le a < b$ and $\mathcal{B}$, the corresponding frequency region $\mathcal{D}$ in (13) can be critically sampled by an integer matrix $\boldsymbol{M}$. It then follows from (14) in Lemma 3 that

$$\widehat{\mathbb{1}}_{\mathcal{D}}(0, \ldots, 0, |\boldsymbol{M}|\,n) = 0, \quad \text{for all } n \in \mathbb{Z} \setminus \{0\}. \quad (15)$$

From the definition of $\mathcal{D}$, we have

$$\begin{aligned}
&\widehat{\mathbb{1}}_{\mathcal{D}}(0, \ldots, 0, x) \\
&= \int_{a \le |\omega_N| \le b} d\omega_N \left(e^{-2\pi j\, x\, \omega_N} \int_{\omega_N \mathcal{B}} 1\; d\omega_1 \ldots d\omega_{N-1} \right) \\
&= \int_{a \le |\omega| \le b} e^{-2\pi j\, x\, \omega}\, m(\omega\, \mathcal{B})\, d\omega \\
&= \int_{a \le |\omega| \le b} e^{-2\pi j\, x\, \omega} |\omega|^{N-1} m(\mathcal{B})\, d\omega \\
&= 2\, m(\mathcal{B}) \int_a^b \omega^{N-1} \cos(2\pi x\, \omega)\, d\omega.
\end{aligned}$$

After a change of variable, we can now rewrite (15) as $\int_{2\pi|\boldsymbol{M}|a}^{2\pi|\boldsymbol{M}|b} \omega^{N-1} \cos(n\,\omega)\, d\omega = 0$, for all $n \in \mathbb{Z} \setminus \{0\}$, which is impossible when $N \ge 3$ by Appendix C of [7]. ∎

5 Conclusions

By linking the alias-free (and critical) sampling of a given frequency support region with the Fourier transform of the indicator function, we presented two simple yet powerful conditions for checking alias-free sampling and critical sampling. We demonstrated the usefulness of the proposed conditions in the design of multidimensional critically sampled filter banks. As an interesting result, we show that it is impossible to construct a *nonredundant* directional filter bank with a general cone-shaped frequency decomposition, except for the 2-D case.

References:

[1] L. Brandolini, L. Colzani, and G. Travaglini. Average decay of Fourier transforms and integer points in polyhedra. *Ark. Mat.*, 35:253–275, 1997.

[2] P. M. Gruber and C. G. Lekkerkerker. *Geometry of Numbers*. Elsevier Science Publishers, Amsterdam, second edition, 1987.

[3] M. N. Kolountzakis and J. C. Lagarias. Tilings of the line by translates of a function. *Duke Math. J.*, 82(3):653–678, 1996.

[4] H. R. Künsch, E. Agrell, and F. A. Hamprecht. Optimal lattices for sampling. *IEEE Trans. Inf. Theory*, 51(2):634–47, Feb. 2005.

[5] Y. M. Lu and M. N. Do. Finding optimal integral sampling lattices for a given frequency support in multidimensions. In *Proc. IEEE Int. Conf. on Image Proc.*, San Antonio, USA, 2007.

[6] Y. M. Lu and M. N. Do. Multidimensional directional filter banks and surfacelets. *IEEE Trans. Image Process.*, 16(4):918–931, April 2007.

[7] Y. M. Lu, M. N. Do, and R. S. Laugesen. A computable Fourier condition generating alias-free sampling lattices. *IEEE Trans. Signal Process.*, to appear, 2009.

[8] D. P. Peterson and D. Middleton. Sampling and reconstruction of wavenumber-limited functions in N-dimensional Euclidean spaces. *Inform. Contr.*, 5:279–323, 1962.

[9] P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice-Hall, Englewood Cliffs, NJ, 1993.

Analysis of Singularities and Edge Detection using the Shearlet Transform

Glenn Easley [(1)], Kanghui Guo [(2)], and Demetrio Labate [(3)]

(1) System Planning Corporation 1000 Wilson Boulevard, Arlington, VA 22209, USA.
(2) Missouri State University, Springfield, MO 65804, USA.
(3) University of Houston, 651 Phillip G Hoffman, Houston, TX 77204-3008, USA.
geasley@sysplan.com, KanghuiGuo@MissouriState.edu, dlabate@math.uh.edu

Abstract:

The continuous curvelet and shearlet transforms have recently been shown to be much more effective than the traditional wavelet transform in dealing with the set of discontinuities of functions and distributions. In particular, the continuous shearlet transform has the ability to provide a very precise geometrical characterization of general discontinuity curves occurring in images. In this paper, we show that these properties are useful to design improved algorithms for the analysis and detection of edges.

1. Introduction

One of the most useful properties of the wavelet transform is its ability to deal very efficiently with the discontinuities of functions and distributions. Consider, for example, a function f on $\mathbb{R}^2$ which is smooth except for a discontinuity at $x_0 \in \mathbb{R}^2$, and let $\mathcal{W}_\psi f(a,t)$ be the *continuous wavelet transform* of f. This is defined as the mapping

$$\mathcal{W}_\psi f(a,t) = a^{-1} \int_{\mathbb{R}^2} f(x)\, \psi\left(a^{-1}(x-t)\right)\, dx,$$

where $a > 0, t \in \mathbb{R}^2$ and $\psi \in L^2(\mathbb{R}^2)$ is an appropriate well-localized function. Then $\mathcal{W}_\psi f(a,t)$ decays rapidly as $a \to 0$ everywhere, unless t is near x_0 [5]. Hence, the wavelet transform is able to signal the location of the singularity of f through its asymptotic decay at fine scales. It was recently shown that certain "directional" extensions of the wavelet transform have the ability to provide a much finer description of the set of singularities of a function. Namely, the recently introduced curvelet and shearlet transforms are able to identify not only the location of singularities of a function, but also the orientation of discontinuity curves. In particular, using the continuous shearlet transform, one can precisely characterize the geometrical information of general discontinuity curves, including discontinuity curves which contain irregularities such as corner and junction points.

In this paper, we show that one can take advantage of the properties of the shearlet transform to design improved algorithms for the analysis and detection of edges in images. Indeed, multiscale techniques based on wavelets have a history of successful applications in the study of edges. With respect to traditional wavelets, the shearlet framework has the ability to capture directly the information about edge orientation and this is useful to improve the robustness of edge detection algorithms in the presence of noise.

The paper is organized as follows. In Section 2. we recall the definition of the shearlet transform and its main results concerning the analysis of edges. In Section 3. we present some representative numerical experiments of edge detection, comparing the shearlet approach against wavelets and other standard edge detection techniques.

2. The Shearlet Transform

For $a > 0$ $s \in \mathbb{R}$ and $t \in \mathbb{R}^2$, let M_{as} be the matrices

$$M_{as} = \begin{pmatrix} a & -\sqrt{a}s \\ 0 & \sqrt{a} \end{pmatrix}$$

and, corresponding to those, let $\psi_{ast}(x) = |\det M_{as}|^{-\frac{1}{2}} \psi(M_{as}^{-1}(x-t))$, where $\psi \in L^2(\mathbb{R}^2)$. It is useful to notice that $M_{as} = B_s\, A_a$, where $A_a = \begin{pmatrix} a & 0 \\ 0 & \sqrt{a} \end{pmatrix}$ and $B_s = \begin{pmatrix} 1 & -s \\ 0 & 1 \end{pmatrix}$. Hence to each matrix M_{as} are associated two distinct actions: an *anisotropic* dilation produced by the matrix A_a and a *shearing* produced by the non-expansive matrix B_s.

For $f \in L^2(\mathbb{R}^2)$, the *continuous shearlet transform* is defined as the mapping

$$f \to \mathcal{SH}_\psi f(a,s,t) = \langle f, \psi_{ast} \rangle,\ a > 0, s \in \mathbb{R}, t \in \mathbb{R}^2.$$

The generating function ψ is chosen to be a well localized function satisfying appropriate admissibility conditions [7, 4], so that each $f \in L^2(\mathbb{R}^2)$ satisfies the generalized Calderòn reproducing formula:

$$f = \int_{\mathbb{R}^2} \int_{-\infty}^{\infty} \int_0^{\infty} \langle f, \psi_{ast} \rangle\, \psi_{ast}\, \frac{da}{a^3}\, ds\, dt.$$

The significance of the shearlet representation is that any function f is broken up with respect to well-localized analyzing elements defined not only at various scales and locations, as in the traditional multiscale approach, but also at various orientations associated with the shearing parameter s. Figure 1 shows the frequency support of the shearlet analyzing functions $\hat{\psi}_{ast}$ for some values of s and a. Thanks to this directional multiscale decomposition, the continuous shearlet transform is able to precisely capture the geometry of edges through its asymptotic decay at fine

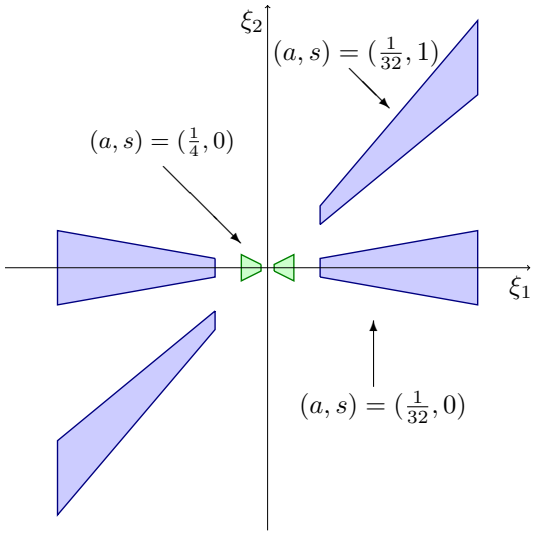


Figure 1: Frequency support of same representative shearlet analyzing functions $\hat{\psi}_{ast}$.

scales ($a \to 0$). To precisely describe these properties, let us introduce the following model of images.
Let $\Omega = [0,1]^2$ and consider the partition $\Omega = \bigcup_{n=1}^{L} \Omega_n \cup \Gamma$, where:

1. each "object" Ω_n, for $n = 1, \ldots, L$, is a connected open set;
2. the set of edges of Ω is given by $\Gamma = \bigcup_{n=1}^{L} \partial_\Omega \Omega_n$, where each boundary $\partial_\Omega \Omega_n$ is a piecewise smooth curve of finite length.

Hence, we consider the space of images $u \in I(\Omega)$ of the form

$$u(x) = \sum_{n=1}^{L} u_n(x)\, \chi_{\Omega_n}(x) \text{ for } x \in \Omega \backslash \Gamma$$

where, for each $n = 1, \ldots, L$, $u_n \in C_0^1(\Omega)$ has bounded partial derivatives, and the sets Ω_n are pairwise disjoint in measure. We have the following result, which is a significant refinement with respect to the simple detection of singularities obtained using traditional wavelets.

Theorem 2.1. *Let $f \in I(\Omega)$.*

(i) If $t \notin \Gamma$, then, for each $N \in \mathbb{N}$

$$\lim_{a \to 0^+} a^{-N}\, \mathcal{SH}_\psi f(a,s,t) = 0.$$

(ii) If $t \in \Gamma$ is a regular point and s does not correspond to the normal direction of Γ at t then

$$\lim_{a \to 0^+} a^{-N}\, \mathcal{SH}_\psi B(a,s,t) = 0, \quad \text{for all } N > 0;$$

otherwise, if $s = s_0$ corresponds to the normal direction of Γ at t then

$$0 < \lim_{a \to 0^+} a^{-\frac{3}{4}} \left|\mathcal{SH}_\psi B(a,s_0,t)\right| < \infty.$$

(iii) If $t \in \Gamma$ is a corner point and s does not correspond to any of the normal directions of Γ at t, then

$$\lim_{a \to 0^+} a^{-\frac{9}{4}} \left|\mathcal{SH}_\psi B(a,s,t)\right| < \infty;$$

otherwise, if $s = s_0$ corresponds to one of the normal directions of Γ at t then

$$0 < \lim_{a \to 0^+} a^{-\frac{3}{4}} \left|\mathcal{SH}_\psi B(a,s_0,t)\right| < \infty.$$

Thus, the continuous shearlet transform has rapid asymptotic decay, as $a \to 0$, everywhere except for locations t on the edges and orientations s which are normal to the edges. We refer to [7, 4, 3] for additional detail, including a more precise description of the behavior at the corner points. We also refer to [1] for some similar (even if more restricted) results based on the curvelet transform.

2.1 Lipschitz regularity

The notion of Lipschitz regularity is a method to quantitatively describe the local regularity of functions and distributions.
Given $\alpha \geq 0$, a function f is Lipschitz α at $x_0 \in \mathbb{R}^2$ if there exists a positive constant K and a polynomial p_{x_0} of degree $m = \lfloor \alpha \rfloor$ such that, for all x in a neighborhood of x_0:

$$|f(x) - p_{x_0}(x)| \leq K\, |x - x_0|^\alpha . \tag{1}$$

A function f is uniformly Lipschitz α over an open set $\Omega \subset \mathbb{R}^2$ if there exists a constant $K > 0$, independent of x_0, such that the above inequality holds for all $x_0 \in \Omega$.
If f is uniformly Lipschitz $\alpha > m$ in a neighborhood of x_0, then f is necessarily m times differentiable at x_0. Also notice that if $0 \leq \alpha < 1$, then $p_{x_0} = f(x_0)$ and condition (1) becomes

$$|f(x) - f(x_0)| \leq K\, |x - x_0|^\alpha .$$

If f is Lipschitz α with $\alpha < 1$ at x_0, then f is not differentiable at x_0. The closer the Lipschitz exponent is to 0, the more "singular" the function is. If f is bounded but discontinuous at x_0, then it is Lipschitz 0 at x_0, indicating the presence of an edge.
Also recall that if $f(x)$ is Lipschitz α, then its primitive $g(x)$ is Lipschitz α + 1 (the converse however is not true; that is, if a function is Lipschitz α at x_0, then its derivative need not be Lipschitz α - 1 at the same point). This observation explains the following definition which extends the concept of Lipschitz regularity to distributions.
Let α be a real number. A tempered distribution f is uniformly Lipschitz α on $\Omega \subset \mathbb{R}^2$ if its primitive is uniformly Lipschitz $\alpha + 1$ on $\Omega \subset \mathbb{R}^2$.
It follows that a distribution may have a negative Lipschitz exponent. For example, one can show that if f is a Dirac delta distribution centered at x_0, then f is Lipschitz -1 at x_0. We refer to [8] and to the references indicated there for more details.
The function ψ satisfies the property that for each $n \in \mathbb{N}$, there exists a constant $c_n > 0$ such that

$$|\psi(x)| \leq c_n (1 + |x|)^{-n}$$

for all $x \in \mathbb{R}^2$ (for details, see [4], p. 26). As a consequence, we obtain $\|\psi\|_1 = \int_{\mathbb{R}^2} |\psi(x)|\, dx < \infty$, and $\int_{\mathbb{R}^2} |\psi(x)||x|^\alpha\, dx < \infty$.
The following result (whose proof is reported in the appendix) is an adaptation of a similar theorem about the

continuous wavelet transform due to Jaffard [6]. If we assume ψ has n vanishing moments, i.e. $\int t^k \psi(t)\, dt = 0$ for all $k = 0, \ldots, n-1$, we would need to add the condition $\alpha \leq n$. However, the general construction of ψ implies that ψ has an infinite number of vanishing moments. Thus this assumption is unnecessary.

Theorem 2.2. *If $f \in L^2(\mathbb{R}^2)$ is Lipschitz $\alpha > 0$ at t_0, then there exists a constant $C > 0$ such that, for all $a < 1$,*

$$|\mathcal{SH}_\psi f(a, s, t)| \leq C\, a^{\frac{1}{2}(\alpha+\frac{3}{2})} \left(1 + \left|a^{-\frac{1}{2}}(t - t_0)\right|\right).$$

The theorem can be extended to the case where f is a distribution. In addition, the estimation of the decay of the shearlet transform of the Dirac delta and other distributions was computed in [7]. These results show that, for locations t corresponding to delta-type singularities, the shearlet transform has a very different behavior from edge points. In fact, the amplitude of $|\mathcal{SH}_\psi f(a, s, t_0)|$ grows like $O(a^{-\frac{1}{4}})$ as $a \to 0$. Similarly, for spike singularities, one can show that the amplitude of the shearlet transform increases at fine scales. This shows that classification of points by their Lipschitz regularity is important as it can be used to distinguish true edge points from points corresponding to noise. This principle was already exploited, for example, in [8].

3. Shearlet-based Edge Detection

Taking advantage of the theoretical observations reported above, a discrete version of the shearlet transform was developed and applied to the purpose of locating and identifying edges in images. Because of space limitations, we will limit ourselves to presenting a few numerical demonstrations. A detailed account of the discrete shearlet transform and shearlet-based edge detection algorithms is found in [2, 10].

Figures 2 and 3 compare a shearlet-based edge detection routine against a wavelet-based routine using a consistent set of predetermined default parameters. For a base-line comparison against standard routines, we also used the Sobel and Prewitt methods using their default parameters. The results highlight the superior performance of the shearlet-based method. To assess the performance of the edge detector, we have given the value of the Pratt's Figure of Merit (FOM), which is a fidelity measure ranging from 0 to 1, with 1 indicating a perfect edge detector [9].

Acknowledgments DL acknowledges partial support from NSF DMS 0604561 and DMS (Career) 0746778.

4. Appendix: Proof of Theorem 2.2.

Proof of Theorem 2.2. Since f is Lipschitz α at t_0, there is a polynomial $p_{t_0}(x)$ and a constant $K > 0$ such that

$$|f(x) - p_{t_0}(x)| \leq K\, |x - t_0|^\alpha.$$

Since $\mathcal{SH}_\psi p_{t_0}(a, s, t) = 0$, then

$$\begin{aligned}
& |\mathcal{SH}_\psi f(a, s, t)| \\
\leq\ & a^{-3/4} \int_{\mathbb{R}^2} |\psi(A_a^{-1} B_s^{-1}(x - t))|\, |f(x) - p_{t_0}(x)|\, dx \\
\leq\ & K\, a^{-3/4} \int_{\mathbb{R}^2} |\psi(A_a^{-1} B_s^{-1}(x - t))|\, |x - t_0|^\alpha\, dx \\
=\ & K\, a^{3/4} \int_{\mathbb{R}^2} |\psi(y)|\, |t + B_s A_a y - t_0|^\alpha\, dy \\
\leq\ & K\, 2^\alpha\, a^{3/4} \left(\|B_s\|^\alpha\, \|A_a\|^\alpha \int_{\mathbb{R}^2} |\psi(y)|\, |y|^\alpha\, dy \right. \\
& \left. + \int_{\mathbb{R}^2} |\psi(y)|\, |t - t_0|^\alpha\, dy \right) \\
\leq\ & K\, 2^\alpha\, a^{3/4} \left(C(s)^\alpha\, a^{\alpha/2} \int_{\mathbb{R}^2} |\psi(y)|\, |y|^\alpha\, dy \right. \\
& \left. + |t - t_0|^\alpha \int_{\mathbb{R}^2} |\psi(y)|\, dy \right) \\
\leq\ & C\, a^{\frac{1}{2}(\alpha+\frac{3}{2})} \left(1 + |a^{-1/2}(t - t_0)|^\alpha\right).
\end{aligned}$$

Here we have used the fact that $\|A_a\| = a^{1/2}$, i.e. the largest eigenvalue of the matrix A_a. Similarly $\|B_s\|$ is the largest eigenvalue of the matrix B_s, which is 1.

References:

[1] E. J. Candès and D. L. Donoho, "Continuous curvelet transform: I. Resolution of the wavefront set", *Appl. Comput. Harmon. Anal.*, Vol. 19, pp. 162–197, 2005. 162–197.

[2] G. Easley, D. Labate, and W-Q. Lim "Sparse Directional Image Representations using the Discrete Shearlet Transform", *Appl. Comput. Harmon. Anal.* Vol. 25, pp. 25–46, 2008.

[3] K. Guo and D. Labate, "Characterization and analysis of edges using the Continuous Shearlet Transform", preprint, 2008

[4] K. Guo, D. Labate and W. Lim, "Edge analysis and identification using the continuous shearlet transform", to appear in *Appl. Comput. Harmon. Anal.*.

[5] M. Holschneider, *Wavelets. Analysis tool*, Oxford University Press, Oxford, 1995.

[6] S. Jaffard "Pointwise smoothness, two-localization and wavelet coefficients", *Publicacions Mathematique*, Vol. 35, pp. 155–168, 1991.

[7] G. Kutyniok and D. Labate, "Resolution of the Wavefront Set using Continuous Shearlets", *Trans. Am. Math. Soc.*, Vol. 361 pp. 2719-2754, 2009.

[8] S. Mallat and W. L. Hwang, Singularity detection and processing with wavelets, *IEEE Trans. Inf. Theory*, vol. 38, no. 2, 617-643, Mar. 1992.

[9] W.K. Pratt, *Digital Image Processing*, Wiley Interscience Publications, 1978.

[10] S. Yi, D. Labate, G. R. Easley, and H. Krim, "A Shearlet Approach to Edge Analysis and Detection", to appear in IEEE Trans. Image processing, 2008.

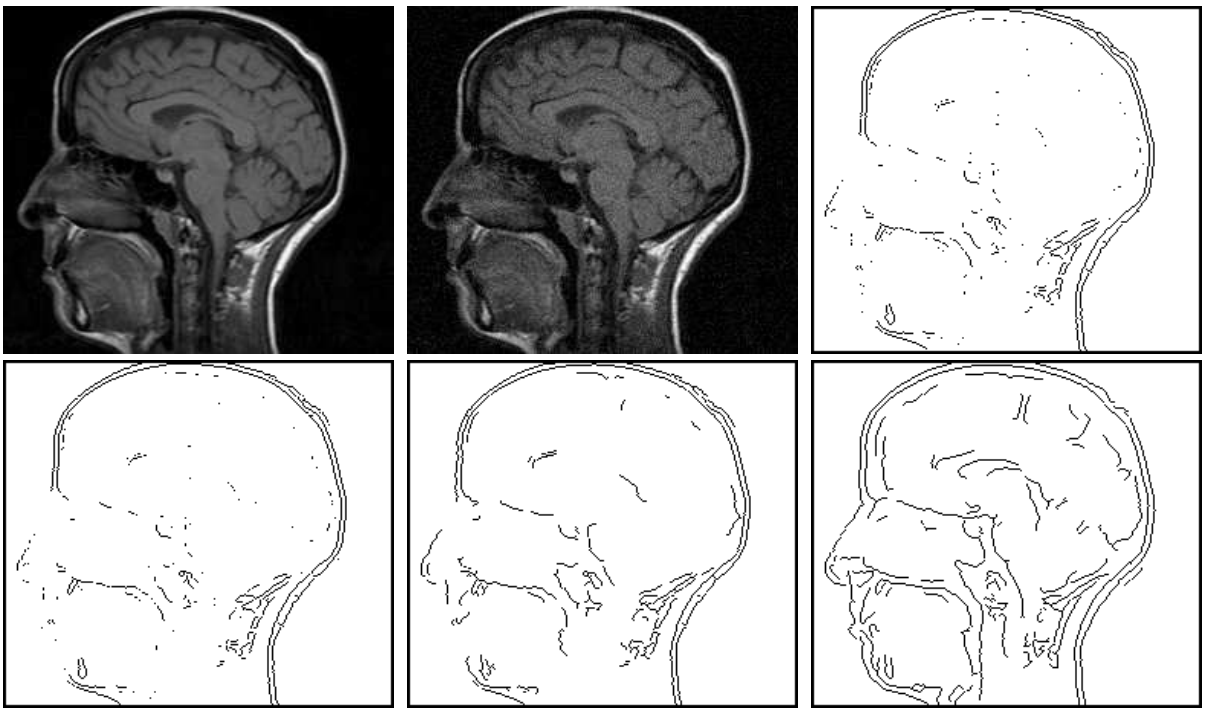


Figure 2: Results of edge detection methods. From top left, clockwise: Original image, noisy image (PSNR=28.10 dB), Sobel result (FOM=0.24), shearlet result (FOM=0.44), wavelet result (FOM=0.29), and Prewitt result (FOM=0.23).

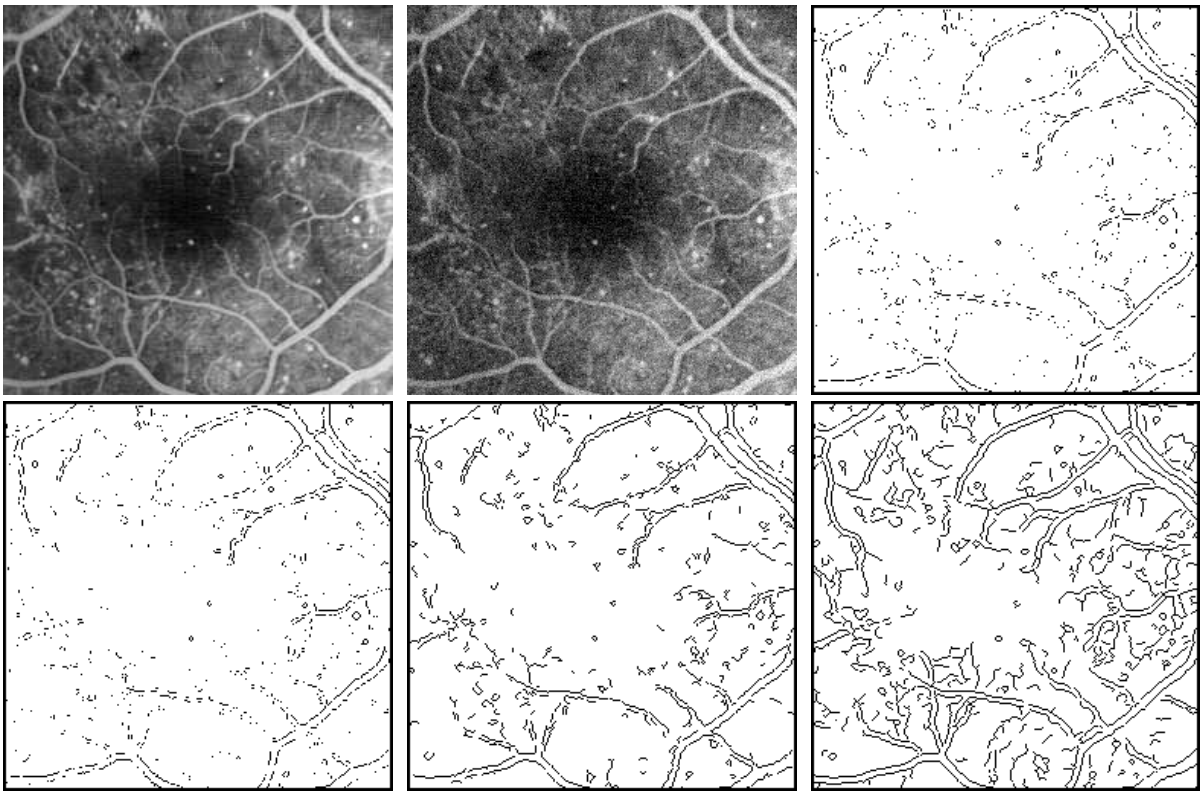


Figure 3: Results of edge detection methods. From top left, clockwise: Original image, noisy image (PSNR=24.58 dB), Sobel result (FOM=0.15), shearlet result (FOM=0.45), wavelet result (FOM=0.27), and Prewitt result (FOM=0.15).

Discrete Shearlet Transform : New Multiscale Directional Image Representation

Wang-Q Lim

Department of Mathematics, University of Osnabrück, Osnabrück, Germany
wlim@mathematik.uni-osnabrueck.de

Abstract:

It is now widely acknowledged that analyzing the intrinsic geometrical features of an underlying image is essentially needed in image processing. In order to achieve this, several directional image representation schemes have been proposed. In this report, we develop the discrete shearlet transform (DST) which provides efficient multiscale directional representation. We also show that the implementation of the transform is built in the discrete framework based on a multiresolution analysis. We further assess the performance of the DST in image denoising and approximation applications. In image approximation, our adaptive approximation scheme using the DST significantly outperforms the wavelet transform (up to 3.0dB) and other competing transforms. Also, in image denoising, the DST compares favorably with other existing methods in the literature.

1. Introduction

Sharp image transitions or singularities such as edges are expensive to represent and intergrating the geometric regularity in the image representation is a key challenge to improve state of the art applications to image compression and denoising. To exploit the anisotropic regularity of a surface along edges, the basis must include elongated functions that are nearly parallel to the edges.

Several image representations have been proposed to capture geometric image regularity. They include curvelets [1], contourlets [2] and bandelets [3]. In particular, the construction of curvelets is not built directly in the discrete domain and they do not provide a multiresolution representation of the geometry. In consequence, the implementation and the mathematical analysis are more involved and less efficient. Contourlets are bases constructed with elongated basis functions using a combination of a multiscale and a directional filter bank. However, contourlets have less clear directional features than curvelets, which leads to artifacts in denoising and compression. Bandelets are bases adapted to the function that is represented. Asymptotically, the resulting bandelets are regular functions with compact support, which is not the case for contourlets. However, in order to find bases adapted to an image, the bandelet transform searches for the optimal geometry. For an image of N pixels, the complexity of this best bandelet basis algorithm is $O(N^{3/2})$ which requires extensive computation [3].

Recently, a new representation scheme has been introduced [4]. These so called *shearlets* are frame elements which yield (nearly) optimally sparse representations [5]. This new representation system is based on a simple and rigorous mathematical framework which not only provides a more flexible theoretical tool for the geometric representation of multidimensional data, but is also more natural for implementations. As a result, the shearlet approach can be associated to a multiresolution analysis [4]. However constructions proposed in [4] do not provide compactly supported shearlets and this property is essentially needed especially in image processing applications. In fact, in order to capture local singularities in images efficiently, basis functions need to be well localized in the spatial domain.

In this report, we construct compactly supported shearlets and show that there is a multiresolution analysis associated with this construction. Based on this, we develop the fast discrete shearlet transform (DST) which provides efficient directional representations.

2. Shearlets

A family of vectors $\{\varphi_n\}_{n\in\Gamma}$ constitutes a *frame* for a Hilbert space $\mathcal{H}$ if there exist two positive constants A, B such that for each $f \in \mathcal{H}$ we have

$$A\|f\|^2 \leq \sum_{n\in\Gamma} |\langle f, \varphi_n\rangle|^2 \leq B\|f\|^2.$$

In the event that $A = B$, the frame is said to be *tight*.

Let us next introduce some notations that we will use throughout this paper. For $f \in L^2(\mathbb{R}^d)$, the Fourier transform of f is defined by

$$\hat{f}(\omega) = \int_{\mathbb{R}^d} f(x) e^{-2\pi i x\cdot\omega} dx.$$

Also, for $t \in \mathbb{R}^d$ and $A \in GL_d(\mathbb{R})$, we define the following unitary operators:

$$T_t(f)(x) = f(x - t)$$

and

$$D_A(f)(x) = |A|^{-\frac{1}{2}} f(A^{-1}x).$$

Finally, for $q \in (\frac{1}{2}, 1]$ and $a > 1$, we define

$$A_0 = \begin{pmatrix} a^q & 0 \\ 0 & a^{\frac{1}{2}} \end{pmatrix} \text{ and } B_0 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

and

$$A_1 = \begin{pmatrix} a^{\frac{1}{2}} & 0 \\ 0 & a^q \end{pmatrix} \text{ and } B_1 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}. \qquad (2)$$

We are now ready to define a shearlet frame as follows. For $c \in \mathbb{R}^+$, $\psi_0^1, \ldots, \psi_0^L, \psi_1^1, \ldots, \psi_1^L \in L^2(\mathbb{R}^2)$ and $\phi \in L^2(\mathbb{R}^2)$, we define

$$\Psi_c^0 = \{\psi_{jkm}^{i,0} \,:\, j,k \in \mathbb{Z},\, m \in \mathbb{Z}^2,\, i = 1, \ldots, L\},$$

$$\Psi_c^1 = \{\psi_{jkm}^{i,1} \,:\, j,k \in \mathbb{Z},\, m \in \mathbb{Z}^2,\, i = 1, \ldots, L\},$$

and

$$\Psi_c^2 = \{T_{cm}\phi \,:\, m \in \mathbb{Z}^2\}$$
$$\cup\{\psi_{jkm}^{i,0} \,:\, j \geq 0, -2^j \leq k \leq 2^j,\, m \in \mathbb{Z}^2,\, i = 1, \ldots, L\}$$
$$\cup\{\psi_{jkm}^{i,1} \,:\, j \geq 0, -2^j \leq k \leq 2^j,\, m \in \mathbb{Z}^2,\, i = 1, \ldots, L\}$$

where

$$\psi_{jkm}^{i,\ell} = D_{A_\ell^{-j} B_\ell^{-k}} T_{cm} \psi_\ell^i \qquad (3)$$

for $\ell = 0, 1$, $m \in \mathbb{Z}^2$, $i = 1, \ldots, L$ and $j, k \in \mathbb{Z}$. If Ψ_c^p is a frame for $L^2(\mathbb{R}^2)$, then we call the functions $\psi_{jkm}^{i,\ell}$ in the system Ψ_c^p *shearlets*.

Observe that each element $\psi_{jkm}^{i,\ell}$ in Ψ_c^p is obtained by applying an anisotropic scaling matrix A_ℓ and a shear matrix B_ℓ to fixed generating functions ψ_ℓ^i. This implies that the system Ψ_c^p can provide window functions which can be elongated along arbitrary directions. Therefore, the geometrical structures of singularities in images can be efficiently represented and analyzed using those window functions. In fact, it was shown that 2-dimensional piecewise smooth functions with C^2-singularities can be approximated with nearly optimal approximation rate using shearlets. We refer to [5] for details. Furthermore, one can show that shearlets can completely analyze the singular structures of piecewise smooth images [6]. In fact, this property of shearlets is useful especially in signal and image processing, since singularities and irregular structures carry essential information in a signal. For example, discontinuities in the intensity of an image indicate the presence of edges. Figure 1 displays examples of shearlets which can be elongated along arbitrary direction in the spatial domain.

3. Construction of Shearlets

In this section, we will introduce some useful sufficient conditions to construct compactly supported shearlets. Using these conditions, we will show that the system Ψ_c^p can be generated by simple separable functions associated with a multiresolution analysis. Furthermore, this leads to the fast DST, and we will discuss this in the next section.

We first discuss sufficient conditions for the existence of compactly supported shearlets. For this, let $\alpha > \max(1, (1-p)\gamma)$ and $\gamma > \max\Big(\frac{\alpha+1}{p}, \frac{1}{1-p}\Big)$ be fixed positive numbers for $0 < p < 1$. We choose $\alpha', \gamma' > 0$ such that $\alpha' \geq \alpha + \gamma$ and $\gamma' \geq \alpha' - \alpha + \gamma$. Then we obtain the following results [7].

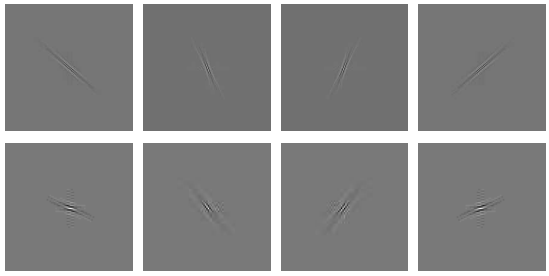


Figure 1: Examples of shearlets in the spatial domain. The top row illustrates shearlet functions $\psi_{jk0}^{i,0}$ associated with matrices A_0 and B_0 in (1). The bottom row shows shearlet functions $\psi_{jk0}^{i,1}$ associated with matrices A_1 and B_1 in (2)..

Theorem 3..1. *[7] For $i = 1, \ldots, L$, we define $\psi_0^i(x_1, x_2) = \gamma^i(x_1)\theta(x_2)$ such that*

$$|\hat{\gamma}^i(\omega_1)| \leq K_1 \frac{|\omega_1|^{\alpha'}}{(1 + |\omega_1|^2)^{\gamma'/2}}$$

and

$$|\hat{\theta}(\omega_1)| \leq K_2(1 + |\omega_1|^2)^{-\gamma'/2}.$$

If

$$\operatorname{ess} \inf_{|\omega_1| \leq 1/2} |\hat{\theta}(\omega_1)|^2 \geq K_3 > 0 \qquad (4)$$

and

$$\operatorname{ess} \inf_{a^{-q} \leq |\omega_1| \leq 1} \sum_{i=1}^{L} |\hat{\gamma}^i(\omega_1)|^2 \geq K_4 > 0, \qquad (5)$$

then there exists $c_0 > 0$ such that Ψ_c^0 is a frame for $L^2(\mathbb{R}^2)$ for all $c \leq c_0$.

Observe that the functions $\psi_0^1, \ldots, \psi_0^L$ are separable functions, and the one-dimensional scaling function θ and wavelets γ^i can be chosen with sufficient vanishing moments in this case.

We now show some concrete examples of compactly supported shearlets using Theorem 3.1. Assume that $a = 4$ and $q = 1$ in (1) and (2). Let us consider a box spline [1] of order m defined as follows.

$$\hat{\theta}_m(\omega_1) = \Big(\frac{\sin \pi\omega_1}{\pi\omega_1}\Big)^{m+1} e^{-i\epsilon\omega_1},$$

where $\epsilon = 1$ if m is even, and $\epsilon = 0$ if m is odd. Obviously, we have the following two scaling equation:

$$\hat{\theta}_m(2\omega_1) = m_0(\omega_1)\hat{\theta}_m(\omega_1)$$

and

$$m_0(\omega_1) = (\cos \pi\omega_1)^{m+1} e^{-i\epsilon\pi\omega_1}.$$

Let α' and γ' be positive real numbers as in Theorem 3.1. We now define

$$\hat{\psi}_0^1(\omega) = (i)^\ell \sqrt{2} \Big(\sin \pi\omega_1\Big)^\ell \hat{\theta}_m(\omega_1)\hat{\theta}_m(\omega_2)$$

and

$$\hat{\psi}_0^2(\omega) = (i)^\ell \Big(\sin \frac{\pi\omega_1}{2}\Big)^\ell \hat{\theta}_m(\frac{\omega_1}{2})\hat{\theta}_m(\omega_2),$$

where $\ell \geq \alpha'$ and $m+1 \geq \gamma'$. Then, by Theorem 3.1, ψ_0^1 and ψ_0^2 generate a frame Ψ_c^0 for $c \leq c_0$ with some $c_0 > 0$. There are infinitely many possible choices for ℓ and m. For example, one can choose $\ell = 9$ and $m = 11$.
Define

$$\phi(x_1, x_2) = \theta_m(x_1)\theta_m(x_2),$$

$$\hat{\psi}_1^1(\omega) = (i)^\ell \sqrt{2}\Big(\sin \pi\omega_2\Big)^\ell \hat{\theta}_m(\omega_2)\hat{\theta}_m(\omega_1)$$

and

$$\hat{\psi}_1^2(\omega) = (i)^\ell \Big(\sin \frac{\pi\omega_2}{2}\Big)^\ell \hat{\theta}_m(\frac{\omega_2}{2})\hat{\theta}_m(\omega_1).$$

Then similar arguments show that ψ_1^1 and ψ_1^2 generate a frame Ψ_c^1 for $c \leq c_0$ with some $c_0 > 0$. Furthermore, the functions ϕ, ψ_ℓ^i for $\ell = 0, 1$ and $i = 1, 2$ generate a frame Ψ_c^2 with $c \leq c_0$ for some $c_0 > 0$.

4. Discrete Shearlet Transform

In the previous section, we constructed compactly supported shearlets generated by separable functions associated with a multiresolution analysis. In this section, we will show that this multiresolution analysis leads to the fast DST which computes $\langle f, \psi_{jkm}^{i,\ell} \rangle$. To be more specific, we let $a = 4$ and $q = 1$ in (1) and (2). For notational convenience, we let $n = (n_1, n_2), m = (m_1, m_2), d = (d_1, d_2) \in \mathbb{Z}^2$ and I_2 be a 2 by 2 identity matrix.
Let $\theta \in L^2(\mathbb{R})$ be a compactly supported function such that $\{\theta(\cdot - n_1) : n_1 \in \mathbb{Z}\}$ is an orthonormal sequence and

$$\theta(x_1) = \sum_{n_1 \in \mathbb{Z}} h(n_1)\sqrt{2}\theta(2x_1 - n_1). \tag{6}$$

Define

$$\gamma(x_1) = \sum_{n_1 \in \mathbb{Z}} g(n_1)\sqrt{2}\theta(2x_1 - n_1) \tag{7}$$

such that γ has sufficient vanishing moments and the pair of the filters h and g is a pair of conjugate mirror filters. We assume that γ and θ satisfy decay conditions (4) and (5) in Theorem 3.1. We also define

$$\phi(x_1, x_2) = \theta(x_1)\theta(x_2),$$

$$\psi_\ell^1(x_1, x_2) = \gamma(x_{\ell+1})\theta(x_{2-\ell}) \tag{8}$$

and

$$\psi_\ell^2(x_1, x_2) = 2^{-\frac{1}{2}}\gamma(\frac{x_{\ell+1}}{2})\theta(x_{2-\ell}) \tag{9}$$

for $\ell = 0, 1$. Then Theorem 3.1 can be easily generalized to show that the functions $\psi_0^1, \psi_0^2, \psi_1^1, \psi_1^2$ and ϕ generate a shearlet frame Ψ_c^2 with $c < c_0$ for some $c_0 > 0$.
Let J be a positive odd integer. Based on a multiresolution analysis associated with the two-scale equation (6), we can now easily derive a fast algorithm for computing shearlet coefficients $\langle f, \psi_{jkm}^{i,\ell} \rangle$ for $\ell = 0, 1, j = 1, \ldots, \frac{J-1}{2}$, and $-2^j \leq k \leq 2^j$ as follows.
First, assume that

$$f = \sum_{n \in \mathbb{Z}^2} f_J(n) D_{2^{-J}I_2} T_n \phi \tag{10}$$

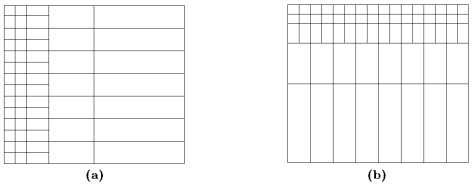


Figure 2: Examples of anisotropic discrete wavelet decomposition: (a) Anisotropic discrete wavelet decomposition by $\mathcal{W}$, (b) Anisotropic discrete wavelet decomposition by $\widetilde{\mathcal{W}}$.

where $f_J(n) = \langle f, D_{2^{-J}I_2} T_n \phi \rangle$. For $h = 0, 1$, let us define maps $\mathcal{D}_h^{k,j} : \ell^2(\mathbb{Z}^2) \to \ell^2(\mathbb{Z}^2)$ by

$$(\mathcal{D}_h^{k,j} x)(d) = \sum_{m \in \mathbb{Z}^2} d_h^{k,j}(d, m)x(m)$$

where $d_h^{k,j}(d, m) = \langle D_{B_h^{k/2^j}} T_m \phi, T_d \phi \rangle$ and $x \in \ell(\mathbb{Z}^2)$. Also we define

$$H(\omega_1) = \sum_{n_1} h(n_1)e^{-2i\pi\omega_1}$$

and

$$G(\omega_1) = \sum_{n_1} g(n_1)e^{-2i\pi\omega_1}.$$

Finally, we let h_j, g_j^0 and g_j^1 be the Fourier coefficients of

$$\begin{cases} H_j(\omega_2) = \prod_{k=0}^{J-j-1} H\Big(2^k\omega_2\Big) \quad \text{for } J - j > 0, \\ G_j^0(\omega_1) = \prod_{k=0}^{J-2j-2} H(2^k\omega_1)G(2^{J-2j-1}\omega_1), \\ G_j^1(\omega_1) = \prod_{k=0}^{J-2j-1} H(2^k\omega_1)G(2^{J-2j}\omega_1), \end{cases} \tag{11}$$

respectively. Then we obtain

$$\begin{cases} \langle f, \psi_{jkm}^{1,0} \rangle = (((\mathcal{D}_0^{k,j} f_J) *_r \overline{h}_j)_{\downarrow 2^{J-j}} *_c \overline{g}_j^0)_{\downarrow 2^{J-2j}}(m), \\ \langle f, \psi_{jkm}^{2,0} \rangle = (((\mathcal{D}_0^{k,j} f_J) *_r \overline{h}_j)_{\downarrow 2^{J-j}} *_c \overline{g}_j^1)_{\downarrow 2^{J-2j+1}}(m), \\ \langle f, \psi_{jkm}^{1,1} \rangle = (((\mathcal{D}_1^{k,j} f_J) *_c \overline{h}_j)_{\downarrow 2^{J-j}} *_r \overline{g}_j^0)_{\downarrow 2^{J-2j}}(m), \\ \langle f, \psi_{jkm}^{2,1} \rangle = (((\mathcal{D}_1^{k,j} f_J) *_c \overline{h}_j)_{\downarrow 2^{J-j}} *_r \overline{g}_j^1)_{\downarrow 2^{J-2j+1}}(m), \end{cases} \tag{12}$$

where $*_c$ and $*_r$ are convolutions along the vertical and horizontal axes respectively, $\downarrow 2^j$ is the downsampling by 2^j and $\overline{h}(n) = h(-n)$ for given filter coefficients $h(n)$.
From (12), we observe that the shearlet transform $\langle f, \psi_{jkm}^{i,\ell} \rangle$ is the application of the shear transformation $D_{B_\ell^{k/2^j}}$ to $f \in L^2(\mathbb{R}^2)$ followed by the wavelet transform associated with anisotropic scaling matrix A_ℓ. In this case, applying $\mathcal{D}_\ell^{k,j}$ to $f_J \in \ell^2(\mathbb{Z}^2)$ corresponds to applying the shear transform $D_{B_\ell^{k/2^j}}$ in the discrete domain. Thus we simply replace the operator $\mathcal{D}_\ell^{k,j}$ by the discrete shear transform $P_{k,j}^\ell$ for $f_J \in \ell^2(\mathbb{Z}^2)$, where we define the discrete shear transforms $P_{k,j}^0$ and $P_{k,j}^1$ as follows:

$$\begin{cases} (P_{k,j}^0 f_J)(n) = f_J\big(n_1 + \lfloor (k/2^j)n_2 \rfloor, n_2\big), \\ (P_{k,j}^1 f_J)(n) = f_J\big(n_1, n_2 + \lfloor (k/2^j)n_1 \rfloor\big). \end{cases} \tag{13}$$

Let M be a fixed positive integer. Since $P_{k,j}^0$ and $P_{k,j}^1$ are unitary operators on $\ell(\mathbb{Z}^2)$, we can extend the shearlet

transform defined in (12) to a linear transform S consisting of finitely many orthogonal transforms S_k^M and $\tilde{S}_k^M$ where

$$S_k^M(f_J) = \mathcal{W}P_{k,M}^0(f_J) \quad \text{and} \quad \tilde{S}_k^M(f_J) = \widetilde{\mathcal{W}}P_{k,M}^1(f_J)$$

and $\mathcal{W}$ and $\widetilde{\mathcal{W}}$ are the wavelet transform associated with an anisotropic sampling matrices A_0 and A_1, respectively. For the precise definitions of $\mathcal{W}$ and $\widetilde{\mathcal{W}}$, we refer to [7]. In this case, the linear transform S, which we call DST, is defined by

$$S = (S_{-2^M}^M, \ldots, S_{2^M}^M, \tilde{S}_{-2^M}^M, \ldots, \tilde{S}_{2^M}^M)$$

for a given $M \in \mathbb{Z}^+$. Notice that redundancy of the DST is $K = 2^{M+2} + 2$ and the DST merely requires $O(KN)$ operations for an image of N pixels. It is obvious that the inverse DST is simply the adjoint of S with normalization.

5. Image Approximation Using DST

In this section, we present some results of the DST in image compression applications. In this case, we use adaptive image representation using the DST. The main idea of this is similar to the matching pursuit introduced by Mallat and Zhong [8]. The matching pursuit selects vectors one by one from a given basis dictionary at each iteration step. On the other hand, our approximation scheme searches the optimal directional index k_0 at each iteration step so that corresponding the orthogonal transform $S_{k_0}^M$ or $\tilde{S}_{k_0}^M$ provides an optimal nonlinear approximation with P nonzero terms among all possible $2^{M+2}+2$ orthogonal transforms in S. For a detailed description of this algorithm, we refer to [7]. For numerical tests, we compare the performance of the DST to other transforms such as the discrete biorthogonal CDF 9/7 wavelet transform (DWT)[9] and contourlet transform (CT)[2] in image compression (see Figure 3). We used only 2 directions (horizontal and vertical) and 4 level decomposition for our DST. In this case, our numerical tests indicate that only a few iterations (1-5) can give significant improvement over other transforms and computing time is comparable to the wavelet transform. For more results, we refer to [8].

6. Conclusion

We have constructed compactly supported shearlet systems which can provide efficient directional image representations. We also have developed the fast discrete implementation of shearlets called the DST. This algorithm consists of applying the shear transforms in the discrete domain followed by the anisotropic wavelet transforms.
Applications of our proposed transform in image approximation and denoising were studied. In image approximation, the results obtained with our adaptive image representation using the DST are significantly superior to those of other transforms such as the DWT and CT both visually and with respect to PSNR.
In denoising, we studied the performance of the DST coupled with a (partially) translation invariant hard tresholding estimator. Our results indicate that the DST consistently outperforms other competing transforms. For detailed numerical results, we refer to [7].

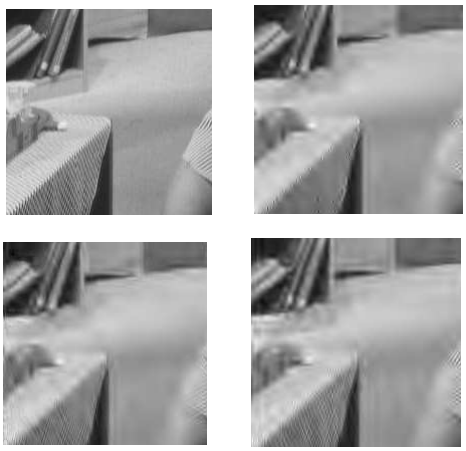


Figure 3: Compression results of 'Barbara' image of size 512×512: The image is reconstructed from 5024 most significant coefficients. **Top left:** Zoomed original image, **Top right:** Zoomed image reconstructed by the DWT (PSNR = 25.11), **Bottom left:** Zoomed image reconstructed by the CT (PSNR = 25.88), **Bottom right:** Zoomed image reconstructed by the DST with only 1 iteration step (PSNR = 26.73).

References:

[1] E. Candes and D. Donoho, "New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities," *Commun. Pure Appl. Math*, vol. 57, no. 2, pp. 219-266, Feb. 2004.

[2] M. Do and M. Vetterli, "The contourlet transform: An efficient directional multiresolution image representation," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2091-2106, Dec. 2005.

[3] G. Peyre and S. Mallat, "Discrete Bandelets with Geometric Orthogonal Filters," *Proceedings of ICIP,* Sept. 2005.

[4] D. Labate, W. Lim, G. Kutyniok and G. Weiss "Sparse Multidimensional Representation using Shearlets", *Proc. of SPIE conference on Wavelet Applications in Signal and Image Processing XI*, San Diego, USA, 2005.

[5] K. Guo and D. Labate, "Optimally Sparse Multidimensional Representation using Shearlets,", *SIAM J Math. Anal.*, 39 pp. 298-318, 2007.

[6] K. Guo, D. Labate and W. Lim, "Edge Analysis and identification using the Continuous Shearlet Transform", to appear in *Appl. Comput. Harmon. Anal.*

[7] W. Lim, "Compactly Supported Shearlet Frames and Their Applications", submitted.

[8] S. Mallat and S. Zhang, "Matching Pursuits With Time-Frequency Dictionaries," *IEEE Trans. Signal Process.*, pp. 3397-3415, Dec. 1993.

[9] A. Cohen, I. Daubechies and J. Feauveau, "Biorthogonal bases of compactly supported wavelets," *Commun. on Pure and Appl. Math.*, 45:485-560, 1992.

Image Approximation by Adaptive Tetrolet Transform

Jens Krommweh

Department of Mathematics, University of Duisburg-Essen, Campus Duisburg, 47048 Duisburg, Germany.
jens.krommweh@uni-due.de

Abstract:

In order to get an efficient image representation we introduce a new adaptive Haar wavelet transform, called **Tetrolet Transform**. Tetrolets are Haar-type wavelets whose supports are tetrominoes which are shapes made by connecting four equal-sized squares. The corresponding filter bank algorithm is simple but enormously effective. Numerical results show the strong efficiency of the tetrolet transform for image compression.

1. Introduction

The main task in every kind of image processing is finding an efficient image representation that characterizes the significant image features in a compact form. In the last years a lot of methods have been proposed to improve the treatment with orientated geometric image structures. Curvelets [1], contourlets [2], shearlets [5], and directionlets [10] are wavelet systems with more directional sensitivity than classical tensor product wavelets.

Instead of choosing a priori a basis or a frame one may adapt the function system depending on the local image structures. Wedgelets [3] and bandelets [7] stand for this second class of image representation schemes which is a wide field of further research. Very recent approaches are the grouplets [8] or the EPWT [9] which are based on an averaging in adaptive neighborhoods of data points.

In [6] we have introduced a new adaptive algorithm whose underlying idea is similar to the idea of digital wedgelets where Haar functions on wedge partitions are considered. We divide the image into 4×4 blocks, then we determine in each block a tetromino partition which is adapted to the image geometry in this block. Tetrominoes are shapes made by connecting four equal-sized squares, each joined together with at least one other square along an edge. On these geometric shapes we define Haar-type wavelets, called *tetrolets*, which form a local orthonormal basis. The main advantage of Haar-type wavelets is the lack of pseudo-Gibbs artifacts. The corresponding filter bank algorithm decomposes an image into a compact representation.

The tetrolet transform is also very efficient for compression of real data arrays.

2. The Adaptive Tetrolet Transform

2.1 Definitions and Notations

Let be $I = \{(i,j) : i,j = 0,\dots,N-1\} \subset \mathbb{Z}^2$ the *index set* of a digital image $\mathbf{a} = (a[i,j])_{(i,j)\in I}$ with $N = 2^J, J \in \mathbb{N}$. We determine a *4-neighborhood* of an index $(i,j) \in I$ by $N_4(i,j) := \{(i-1,j),(i+1,j),(i,j-1),(i,j+1)\}$. An index that lies at the boundary has three neighbors, an index at the vertex of the image has two neighbors.
A set $E = \{I_0,\dots,I_r\}, r \in \mathbb{N}$, of subsets $I_\nu \subset I$ is a *disjoint partition* of I if $I_\nu \cap I_\mu = \emptyset$ for $\nu \neq \mu$ and $\bigcup_{\nu=0}^{r} I_\nu = I$.
In this paper we consider disjoint partitions of the index set I that satisfy two conditions for all I_ν:

1. each subset I_ν contains four indices, i.e. $\#I_\nu = 4$,
2. every index of I_ν has a neighbor in I_ν, i.e. $\forall (i,j) \in I_\nu \; \exists (i',j') \in I_\nu : (i',j') \in N_4(i,j)$.

We call such subsets I_ν *tetromino*, since the tiling problem of the square $[0,N)^2$ by shapes called tetrominoes is a well-known problem being closely related to our partitions of the index set $I = \{0,1,\dots,N-1\}^2$. We shortly introduce this tetromino tiling problem in the next subsection.

2.2 Tilings by Tetrominoes

Tetrominoes were introduced by Golomb in [4]. They are shapes formed from a union of four unit squares, each connected by edges, not merely at their corners. The tiling problem with tetrominoes became popular through the famous computer game classic 'Tetris'. Disregarding rotations and reflections there are five different shapes, the so called *free tetrominoes*, see Figure 1.
It is clear that every square $[0,N)^2$ can be covered by tetrominoes if and only if N is even. But the number of different coverings explodes with increasing N. There are 117 solutions for disjoint covering of a 4×4 board with four tetrominoes. As represented in Figure 2, we have 22

Figure 1: The five free tetrominoes.

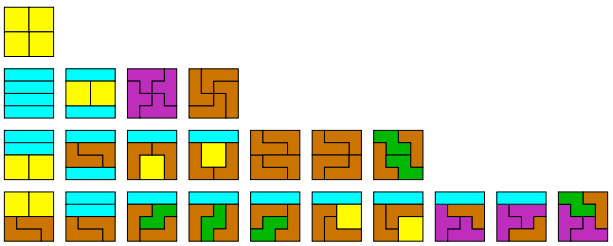


Figure 2: The 22 fundamental forms tiling a 4×4 board. Regarding additionally rotations and reflections there are 117 solutions.

fundamental configurations (disregarding rotations and reflections). One solution (first line) is unaltered by rotations and reflections, four solutions (second line) give a second version applying the isometries. Seven forms can occur in four orientations (third line), and ten asymmetric cases in eight directions (last line).

2.3 The Idea of Tetrolets

In the two-dimensional classical Haar case, the low-pass filter and the high-pass filters are just given by the averaging sum and the averaging differences of each four pixel values which are arranged in a 2×2 square, i.e., with $I_{i,j} = \{(2i, 2j), (2i+1, 2j), (2i, 2j+1), (2i+1, 2j+1)\}$ for $i, j = 0, 1, \ldots, \frac{N}{2} - 1$, we have a dyadic partition $E = \{I_{0,0}, \ldots, I_{\frac{N}{2}-1,\frac{N}{2}-1}\}$ of the image index set I. Let L be a bijective mapping which maps the four pixel pairs (i, j) to the scalar set $\{0, 1, 2, 3\}$, that means it brings the pixels into a unique order. Then we can determine the low-pass part $\mathbf{a}^1 = (a^1[i, j])_{i,j=0}^{\frac{N}{2}-1}$ as well as the three high-pass parts $\mathbf{w}_l^1 = (w_l^1[i, j])_{i,j=0}^{\frac{N}{2}-1}$ for $l = 1, 2, 3$ with

$$a^1[i, j] = \sum_{(i',j') \in I_{i,j}} \epsilon[0, L(i', j')]\, a[i', j'] \quad (1)$$

$$w_l^1[i, j] = \sum_{(i',j') \in I_{i,j}} \epsilon[l, L(i', j')]\, a[i', j'], \quad (2)$$

where the coefficients $\epsilon[l, m], l, m = 0, \ldots, 3$, are entries from the Haar wavelet transform matrix

$$W := (\epsilon[l, m])_{l,m=0}^{3} = \frac{1}{2} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}. \quad (3)$$

Obviously, the fixed blocking by the dyadic squares $I_{i,j}$ is very inefficient because the local structures of an image are disregarded. Our idea is, to allow more general partitions such that the local image geometry is taken into account. Namely, we use tetromino partitions. As described in the previous subsection we shall restrict us to 4×4 blocks. This leads to a third condition for the desired disjoint partition E of the index set I introduced in Section 2.1:

3. Each 4×4 square $Q_{i,j} := \{4i, \ldots, 4i + 3\} \times \{4j, \ldots, 4j + 3\}, i, j = 0, 1, \ldots, \frac{N}{4} - 1$, is covered by four subsets (tetrominoes) $I_0, \ldots, I_3$.

In other words, we first divide the index set I of an image $\mathbf{a}$ into $\frac{N^2}{16}$ squares $Q_{i,j}$ and then we consider the admissible tetromino partitions there. Among the 117 solutions we compute an optimal partition in each image block such that the wavelet coefficients defined on the tetrominoes have minimal l^1-norm.

3. Detailed Description of the Algorithm

The rough structure of the tetrolet filter bank algorithm is described in Table 1.

Adaptive Tetrolet Decomposition Algorithm
Input: Image $\mathbf{a} = (a[i, j])_{i,j=0}^{N-1}$ with $N = 2^J, J \in \mathbb{N}$.
1. Divide the image into 4×4 blocks.
2. Find in each block the sparsest tetrolet representation.
3. Rearrange the low- and high-pass coefficients of each block into a 2×2 block.
4. Store the tetrolet coefficients (high-pass part).
5. Apply step 1 to 4 to the low-pass image.
Output: Decomposed image $\tilde{\mathbf{a}}$.

Table 1: Adaptive tetrolet decomposition algorithm.

Going into detail our main attention shall be turned to step 2 of the algorithm where the adaptivity comes into play. We start with the input image $\mathbf{a}^0 = (a[i, j])_{i,j=0}^{N-1}$ with $N = 2^J, J \in \mathbb{N}$. In the rth-level, $r = 1, \ldots, J - 1$, we apply the following computations.

1. Divide the low-pass image $\mathbf{a}^{r-1}$ into blocks $Q_{i,j}$ of size $4 \times 4, i, j = 0, \ldots, \frac{N}{4^r} - 1$.

2. In each block $Q_{i,j}$ we compute analogously to (1) and (2) the pixel averages for every admissible tetromino covering $c = 1, \ldots, 117$ by

$$a^{r,(c)}[s] = \sum_{(m,n) \in I_s^{(c)}} \epsilon[0, L(m, n)]\, a^{r-1}[m, n],$$

as well as the three high-pass parts for $l = 1, 2, 3$

$$w_l^{r,(c)}[s] = \sum_{(m,n) \in I_s^{(c)}} \epsilon[l, L(m, n)]\, a^{r-1}[m, n],$$

$s = 0, \ldots, 3$, where the coefficients are given in (3) and L is the mapping mentioned above. Then we choose the covering c^* such that the l^1-norm of the tetrolet coefficients becomes minimal

$$c^* = \arg\min_c \sum_{l=1}^{3} \sum_{s=0}^{3} |w_l^{r,(c)}[s]|. \quad (4)$$

Hence, for every block $Q_{i,j}$ we get an optimal tetrolet decomposition $[\mathbf{a}^{r,(c^*)}, \mathbf{w}_1^{r,(c^*)}, \mathbf{w}_2^{r,(c^*)}, \mathbf{w}_3^{r,(c^*)}]$. By doing this, the local structure of the image block is adapted. The best configuration c^* is a covering whose tetrominoes do not intersect an important structure like an edge in the image $\mathbf{a}^{r-1}$. Because the tetrolet coefficients become as minimal as possible a sparse image representation will be obtained. We have to store for each block $Q_{i,j}$ which covering c^* has been chosen, since this information is necessary for reconstruction.

3. In order to be able to apply further levels of the tetrolet decomposition algorithm, we rearrange the entries of the vectors $\mathbf{a}^{r,(c^*)}$ and $\mathbf{w}_l^{r,(c^*)}$ into 2×2 matrices,

$$\mathbf{a}^r_{|_{Q_{i,j}}} = \begin{pmatrix} a^{r,(c^*)}[0] & a^{r,(c^*)}[2] \\ a^{r,(c^*)}[1] & a^{r,(c^*)}[3] \end{pmatrix},$$

and in the same way $\mathbf{w}^r_{l|_{Q_{i,j}}}$ for $l = 1, 2, 3$.

4. After finding a sparse representation in every block $Q_{i,j}$ for $i, j = 0, \ldots, \frac{N}{4^r} - 1$, we store (as usually done) the low-pass matrix $\mathbf{a}^r$ and the high-pass matrices $\mathbf{w}_l^r, l = 1, 2, 3$, replacing the low-pass image $\mathbf{a}^{r-1}$ by the matrix

$$\begin{pmatrix} \mathbf{a}^r & \mathbf{w}_2^r \\ \mathbf{w}_1^r & \mathbf{w}_3^r \end{pmatrix}.$$

After a suitable number of decomposition steps, one can apply a shrinkage to the tetrolet coefficients in order to get a sparse image representation.

4. An Orthonormal Basis of Tetrolets

We describe the discrete basis functions which correspond to the above algorithm. Remember that the digital image $\mathbf{a} = (a[i,j])_{(i,j)\in I}$ is a subset of $l_2(\mathbb{Z}^2)$. For any tetromino I_ν of I we define the discrete functions

$$\phi_{I_\nu}[m,n] := \begin{cases} 1/2, & (m,n) \in I_\nu, \\ 0, & \text{else}, \end{cases}$$

$$\psi^l_{I_\nu}[m,n] := \begin{cases} \epsilon[l, L(m,n)], & (m,n) \in I_\nu, \\ 0, & \text{else}. \end{cases}$$

Due to the underlying tetromino support, we call ϕ_{I_ν} and $\psi^l_{I_\nu}$ *tetrolets*. As a straightforward consequence of the orthogonality of the standard 2D Haar basis functions and the disjoint partition of the discrete space by the tetromino supports, we have the following essential statement.

Theorem 1 *For every admissible covering $\{I_0, I_1, I_2, I_3\}$ of a 4×4 square $Q \subset \mathbb{Z}^2$ the tetrolet system*

$$\{\phi_{I_\nu} : \nu = 0, 1, 2, 3\} \cup \{\psi^l_{I_\nu} : \nu = 0, 1, 2, 3; l = 1, 2, 3\}$$

is an orthonormal basis of $l^2(Q)$.

5. Cost of Adaptivity: Modified Tetrolet Transform

We will address the costs of storing additional adaptivity information. Our observations will lead to some relaxed versions of the tetrolet transform in order to reduce these costs.

It is well known that a vector of length N and with entropy E can be stored with $N \cdot E$ bits. Hence, the entropy describes the required bits per pixel (bpp) and is an appropriate measure for the quality of compression.

In the following, we propose three methods of entropy reduction in order to reduce the adaptivity costs. An application of these modified transforms as well as of combinations of them is given in the last section.

The simplest approach of entropy reduction is reduction of the symbol alphabet. The tetrolet transform uses the alphabet $\{1, \ldots, 117\}$ for the chosen covering in each image block. If we restrict ourselves to 16 essential configurations that feature different directions we considerably reduce the entropy as well as the computation time.

A second approach to reduce the entropy is to change the distribution of the symbols. Relaxing the tetrolet transform we could ensure that only very few tilings are preferred. Hence, we allow the choice of an *almost* optimal covering c^* in (4) in order to get a tiling which is already frequently chosen. More precisely, we replace (4) by the two steps:

1. Find the set of almost optimal configurations that satisfy

$$\sum_{l=1}^{3}\sum_{s=0}^{3} |w_l^{r,(c)}[s]| \leqslant \min_c \sum_{l=1}^{3}\sum_{s=0}^{3} |w_l^{r,(c)}[s]| + \theta$$

with a predetermined tolerance parameter θ.

2. Among these tilings choose the covering c which is chosen most frequently in the previous image blocks.

Using an appropriate relaxing parameter θ, we achieve a satisfactory balance between low entropy (low adaptivity costs) and minimal tetrolet coefficients.

The third method also reduces the entropy by optimization of the tiling distribution. After an application of an edge detector we use the classical Haar wavelet transform inside flat image regions. In the image blocks that contain edges we make use of the strong adaptivity of the proposed tetrolet transform.

More details of the modified versions can be found in [6].

6. Numerical Experiments

We apply a complete wavelet decomposition of an image and use a shrinkage with global hard-thresholding.

The detail 'monarch' image in Figure 3 shows the enormous efficiency in handling with several directional edges due to the high adaptivity. It can be well noticed that the tetrolet transformation gives excellent results for piecewise constant images. Though the tetrolets are not continuous the approximation of the 'cameraman' image in Figure 4 illustrates that even for natural images the tetrolet filter bank outperforms the tensor product wavelets with the biorthogonal 9-7 filter bank, since no pseudo-Gibbs phenomena occur. This confirms the fact already noticed with wedgelets [3] and bandelets [7]: While nonadaptive methods need smooth wavelets for excellent results, well constructed adaptive methods need not. See [6] for more numerical examples.

Considering the adaptivity costs we compare the standard tetrolet transform with its modified versions. Of course, reduction of adaptivity cost produces a loss of approximation quality. Hence, a satisfactory balance is necessary.

For a rough estimation of the complete storage costs of the compressed image with N^2 pixels we apply a simplified scheme

$$cost_{full} = cost_W + cost_P + cost_A,$$

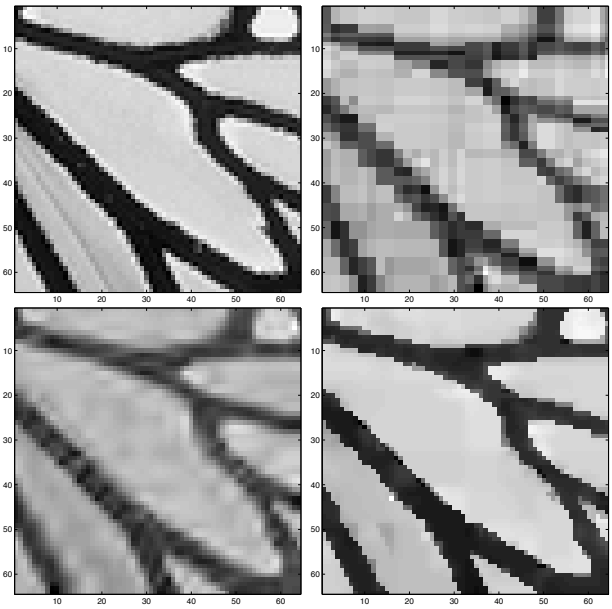


Figure 3: Approximation with 256 coefficients. (a) Input, (b) classical Haar, PSNR 18.98, (c) Biorthogonal 9-7, PSNR 21.78, (d) Tetrolets, PSNR 24.43.

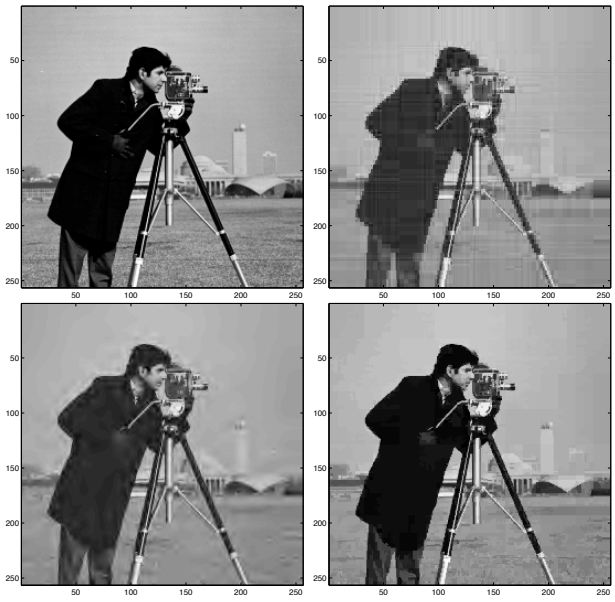


Figure 4: Approximation with 2048 coefficients. (a) Input, (b) classical Haar, PSNR 25.47, (c) Biorthogonal 9-7, PSNR 27.26, (d) Tetrolets, PSNR 29.17.

where $cost_W = 16 \cdot M/N^2$ are the costs in bpp of storing M non-zero wavelet coefficients with 16 bits. The term $cost_P$ gives the cost for coding the position of these M coefficients by $-\frac{M}{N^2}\log_2(\frac{M}{N^2}) - \frac{N^2-M}{N^2}\log_2(\frac{N^2-M}{N^2})$. The third component appearing only with the tetrolet transform contains the cost of adaptivity, $cost_A = E \cdot R/N^2$, for R adaptivity values and the entropy E previously discussed.

Table 2 presents some results for the monarch detail image (Fig. 3) where different versions of the tetrolet transform are compared with the tensor product wavelet transformation regarding to quality and storage costs. We have tried to balance the modified tetrolet transform such that the full costs are in the same scale as with the 9-7 filter. For the relaxed versions we have used the parameter $\theta = 25$.

	coeff	PSNR	entropy	$cost_{full}$
Tensor Haar	300	19.58	-	1.55
Tensor 9-7 filter	300	22.62	-	1.55
Tetrolet	256	24.43	0.53	1.86
Tetro 16	256	23.56	0.30	1.64
Tetro rel	256	24.51	0.32	1.66
Tetro edge	256	24.24	0.43	1.77
Tetro 16 edge rel	256	23.48	0.21	1.55

Table 2: Comparison between tensor wavelet transforms and the different versions of the tetrolet transform regarding quality (PSNR) and storage cost ($cost_{full}$ in bpp).

7. Acknowledgments

The research is funded by the project PL 170/11-1 of the Deutsche Forschungsgemeinschaft (DFG). This is gratefully acknowledged.

References:

[1] E.J. Candes and D.L. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Communications on Pure and Applied Mathematics*, 57(2):219–266, 2004.

[2] M.N. Do and M. Vetterli. The contourlet transform: an efficient directional multiresolution image representation. *IEEE Transactions on Image Processing*, 14(12):2091–2106, 2005.

[3] D.L. Donoho. Wedgelets: Nearly-minimax estimation of edges. *Annals of Statistics*, 27(3):859–897, 1999.

[4] S.W. Golomb. *Polyominoes*. Princeton University Press, 1994.

[5] K. Guo and D. Labate. Optimally sparse multidimensional representation using shearlets. *SIAM Journal on Mathematical Analysis*, 39(1):298–318, 2007.

[6] Jens Krommweh. Tetrolet transform: A new adaptive Haar wavelet algorithm for sparse image representation. 2009.

[7] E. Le Pennec and S. Mallat. Sparse geometric image representations with bandelets. *IEEE Transactions on Image Processing*, 14(4):423–438, 2005.

[8] S. Mallat. Geometrical grouplets. *Applied and Compu-tational Harmonic Analysis*, 26(2):161–180, 2009.

[9] G. Plonka. Easy path wavelet transform: A new adaptive wavelet transform for sparse representation of two-dimensional data. *Multiscale Modeling and Simulation*, 7(3):1474–1496, 2009.

[10] V. Velisavljevic, B. Beferull-Lozano, M. Vetterli, and P.L. Dragotti. Directionlets: Anisotropic multidirectional representation with separable filtering. *IEEE Transactions on Image Processing*, 15(7):1916–1933, 2006.

Geometric Wavelets for Image Processing: Metric Curvature of Wavelets

Emil Saucan [(1)], Chen Sagiv [(2)] and Eli Appleboim [(3)]

(1) Department of Mathematics, Technion - Israel Institute of Technology, Haifa 32000, Israel.
(2) SagivTech Ltd. Israel.
(3) Electrical Engineering Department, Technion - Israel Institute of Technology, Haifa 32000, Israel.
semil@tx.technion.ac.il, chensagivron@gmail.com, eliap@ee.technion.ac.il

Abstract:

We introduce a semi-discrete version of the Finsler-Haantjes metric curvature to define curvature for wavelets and show that scale and curvature play similar roles with respect to image presentation and analysis. More precisely, we show that there is an inverse relationship between local scale and local curvature in images. This allows us to use curvature as a geometrically motivated automatic scale selection in signal and image processing, this being an incipient bridging of the gap between the methods employed in Computer Graphics and Image Processing.

A natural extension to ridgelets and curvelets is also given. Further directions of study, in particular the development of a curvature transform and the study of its link with wavelet and the scale transforms are also suggested.

1. Introduction

The versatility and adaptability of wavelets for a variety of tasks in Image Processing and related fields is too well established in the scientific community, and the bibliography pertaining to it is far too extensive, to even begin to review it here.

We do, however, stress the fact that the multiresolution property of wavelets has been already applied in determining the curvature of planar curves [1] and to the intelligence and reconstruction of meshed surfaces (see, e.g. [18], [26], amongst many others). Moreover, the intimate relation between scale and differentiability in natural images has also been stressed [10].

We have presented in [24] and other related works, an extension of Shannon's Sampling Theorem when images are viewed as higher dimensional objects (i.e. manifolds), rather than 2-dimensional signals. More precisely, our approach to Shannon's Sampling Theorem is based on sampling the graph of the signal, considered as a manifold, rather than sampling of the domain of the signal, as is customary in both theoretical and applied signal and image processing, motivated by the framework of harmonic analysis. The main tool for proving our geometric sampling theorem, resides in the confluence of Differential Topology and Differential Geometry. More precisely, we consider piecewise-linear (PL) approximations of the manifold, where the geometric feature (i.e. curvature) determines the proper size and shape-ration of the simplices of the constructed triangulation.

Naturally, the question is whether the implementation of the geometric sampling scheme is feasible. We do not address here the purely geometric aspects, that would be highly relevant in Computer Graphics implementation (besides, these were partly addressed in [24]). Instead, we focus on the far more important and popular Image Processing tool of wavelets. The versatility and adaptability of wavelets to a variety of tasks in Image Processing and related fields is too well established in the scientific community, and the bibliography pertaining to it is far to extensive, to even begin to review it here.

Unfortunately, in contrast to Computer Graphics experts, for many investigators concerned with wavelets applications, piecewise-linear approximations are not necessarily among their most familiar tools. It is, therefore, a challenge to consider the integration of tools practiced by both communities. Although it may appear to be a surprising result to those primarily familiar with classical wavelets, the *Strömberg wavelets* [27], are based on piecewise-linear functions. Another, more intriguing issue is whether one can replace the intuitive trade-off between scale and curvature, by a formal concept of *wavelet curvature*, in particular in cases such as those of the Strömberg wavelets, or, in the more difficult case of Haar wavelets that are not even piecewise linear.

Interestingly enough, this can be done by using *metric curvatures* [2] (and [21] for a short presentation). It turns out that the best candidate, for the desired metric curvature is the *Finsler-Haantjes curvature*, due to its adaptability to both continuous and discrete settings.

A more suitable approach to surface reconstruction could, for example, implement *ridgelets* [5], or the more generalized, *curvelets* [6].

2. Mathematical Background

The central mathematical concept of the present paper is the following metric notion of curvature suggested by Finsler and developed by Haantjes [12]:

Definition 1 Let (M,d) be a metric space, let $c : I = [0,1] \xrightarrow{\sim} M$ be a homeomorphism, and let $p, q, r \in c(I),\ q, r \neq p$. Denote by $\widehat{qr}$ the arc of $c(I)$ between q and r, and by qr segment from q to r. We say that c has

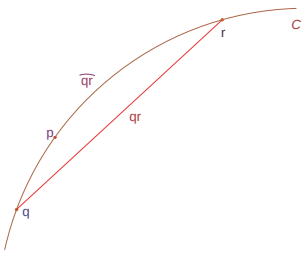


Figure 1: A metric arc and a metric segment.

Finsler-Haantjes curvature $\kappa_{FH}(p)$ at the point p iff:

$$\kappa_{FH}^2(p) = 24 \lim_{q,r \to p} \frac{l(\widehat{qr}) - d(q,r)}{\big(d(q,r)\big)^3} \; ; \qquad (1)$$

where "$l(\widehat{qr})$" denotes the length, in intrinsic metric induced by d, of $\widehat{qr}$ – see Figure 1. (Here we assume that the arc $\widehat{qr}$ has finite length.)

Note that, while highly intuitive and definable for a very large class of curves in general rather metric spaces, this definition of curvature would remain some esoteric notion, without the following theorem (see [2]):

Theorem 2 *Let $c \in \mathcal{C}^3(I)$ be a smooth curve in $\mathbb{R}^3$, and let $p \in c$ be a regular point. Then $\kappa_{FH}(p)$ exists and, moreover, $\kappa_{FH}(p) = k(p)$ – the classical (differential) curvature of c at p.*

3. Finsler-Haantjes Curvature of Wavelets

In [23] we have introduced, in the context of both vertex and edge weighted graphs, a discretization of the Finsler-Haantjes curvature, (for applications in DNA analysis). Here we consider a semi-discrete (or semi-continuous) version, as follows:
Let φ be the typical piecewise-linear wavelet depicted in Figure 2, let $\widehat{AE}$ be the arc of curve between the points A and E, and let $d(A,E)$ is the length of the line-segment AE. Then $l(\widehat{AE}) = a+b+c+d$ and $d(A,E) = e+f$. Then $\kappa_{FH}^2(\varphi) = 24[(a+b+c+d)-(e+f)]/(a+b+c+d)^3$. Note that, in addition to the "total" curvature of φ, one can also compute the "local" curvatures at the "peaks" B and D: $\kappa_{FH}^2(B) = 24(a+c-e)/(a+b)^3$ and $\kappa_{FH}^2(D) = 24(c+d-f)/(a+b)^3$, as well as the mean curvature of these peaks: $\kappa = [\kappa_{FH}(B) + \kappa_{FH}(B)]/2$. Even if these variations may prove to be useful in certain applications, we believe that the correct approach, in the sense that it best corresponds to the scale of the wavelet, would be to compute the total curvature of φ.
Let us compare the relationship between curvature and scale, for a concrete piecewise-linear wavelet – the *Meyer wavelet* [19] – see Figure 3. The results indicating the relationship between scale and curvature, for this wavelet, can be seen in the graph in Figure 4.
However, had the definition of Finsler-Haantjes curvature been limited solely to piecewise-linear wavelets, its applicability would have also been diminished. We show,

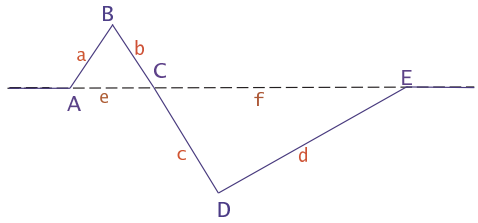


Figure 2: A piecewise-linear wavelet.

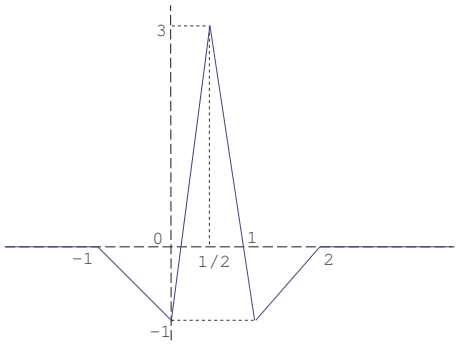


Figure 3: The Meyer wavelet.

however, that it is also definable for the "classical" Haar wavelets, in a rather straightforward manner. For example, consider the basic Haar wavelet and Haar scaling function, illustrated in Figure 5. Then for the scaling function we have: $l(\widehat{AE}) = d(A,B) + d(B,C) + d(C,D) = 3$, while $d(A,D) = 1$. Analogously, for the Haar wavelet we get: $l(\widehat{AE}) = d(M,N,) + d(N,P) + d(P,R) + d(R,S) + d(S,T) = 5$ and $d(M,T) = 1$. The expression for κ_{HF} follow easily in both cases and we present the results for the first 10 scales in Figure 6 and Figure 7, respectively.
Moreover, while perhaps of lesser interest, it should be mentioned that $\kappa_{HF}(\varphi)$ can also be computed for smooth wavelets, using the classical formula for the arc-length: $l(\widehat{AE}) = \int_{\mathrm{Supp}\varphi} \sqrt{1+(\varphi')^2}$.

4. Ridgelets and beyond

The wavelet curvature definition introduced above is applicable, through standard methods, for image processing goals, by using separable 2-dimensional wavelets. However, while practical in many cases, this presumption contravenes to real geometric structure of images, as emphasized, for instance, in [24]. In addition, as it has already been pointed out by Candès [5], "that wavelets can efficiently represent only a small range of the full diversity of interesting behavior", since wavelets can cope well with pointlike singularities, but they are not fitted for the analysis and reconstruction of singularities of dimension greater that 0, that are distributed along lines (and more general curves), planes (and other surfaces), etc. It is therefore natural to ask whether the notion of curvature defined for wavelets can be extended to ridgelets as well.
The perhaps somewhat surprising answer is that such an extension is not only possible, it is in fact more straight-

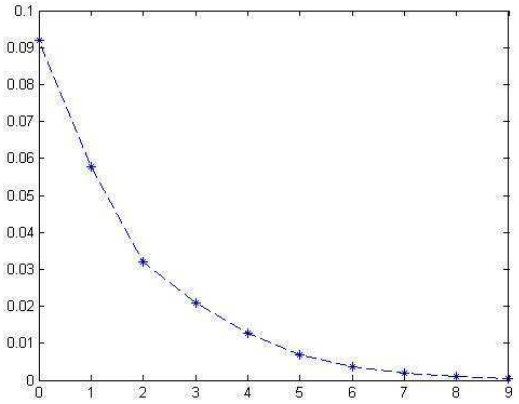


Figure 4: Curvature as a function of scale: Meyer wavelets.

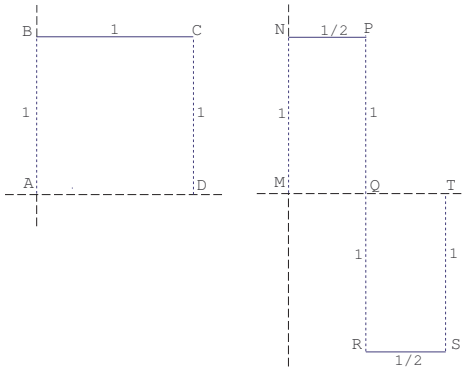


Figure 5: The Harr scaling function and wavelet.

forward and canonical. Indeed, 2-dimensional ridgelets are, in fact, piecewise $\mathcal{C}^2$ surfaces (with line singularities). For these geometrical objects an almost standard notion of curvature exists: the *principal curvatures* (i.e maximal and minimal *normal sectional curvatures* – see [8]) at any point of the surfaces. For ridgelets, we consider only the maximal absolute curvature at points on the ridges (since, along the ridge-line, curvature is 0 (cf. [8]) – see Figure 8. The sectional curvature of curves normal to the ridge is then computed using the method described in the previous section. (See also [22] for the application of the this method to piecewise-flat surfaces.)

Note that similar consideration apply with regard to curvelets (and, evidently, to nonseparable 2-dimensional wavelets as well). However, as far as curvature is concerned, there exists a basic difference between curvelets and ridgelets, which is a direct consequence of the difference between the geometric models employed. Namely, as already noted above, the principal curvature associated with the feature of interest (i.e. the ridge) vanishes. In consequence, Gaussian curvature, being the product of the principal curvatures, will also equal 0 for any point on the ridge (see Figure 8). In contrast, curvelets, being modeled on more flexible types of surfaces, can – and will – exhibit Gaussian curvatures different from 0, both positive and negative.

This geometric analysis can also be applied to shear-

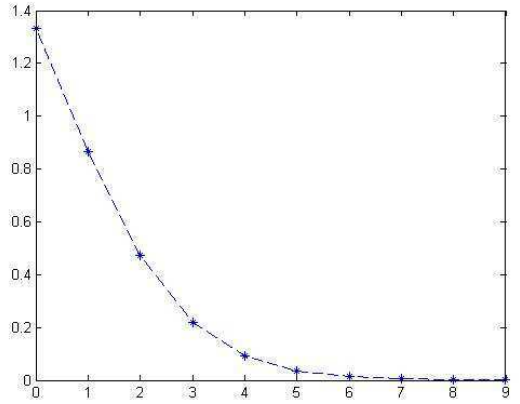


Figure 6: Curvature as a function of scale: The Haar scaling functions.

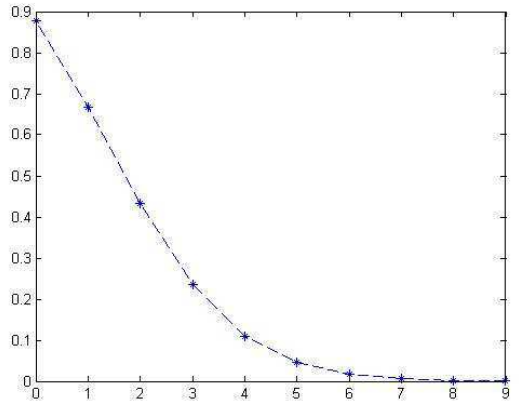


Figure 7: Curvature as a function of scale: The Haar wavelets.

lets. As Figure 9 illustrates, shearlets display "peaks" of high positive Gauss curvature. In consequence, they are ideally suited for modeling phenomena which, in geometric terms, are characterized by positive curvature concentrated at specific points. In view of this, shearlets may be viewed, in the context of our geometric approach, as a complementary tool to ridgelets. Indeed, recall that ridgelets were developed as an extension of wavelets, befitting the modeling of line-type singularities. Point type singularities can still occur in conjunction to 1-dimensional singularities (not least as noise), hence a combination of both type of tools, in a common, integrated "dictionary" is, indeed, required. The geometric approach presented above enables us to build such a "dictionary" in natural manner.

5. Future work – Theory and Applications

As we have seen, curvature can serve as a local scale estimator that is natural, i.e. intrinsic to the geometry of the image. Moreover, it can be easily calculated and used for image analysis and enhancement, especially in edge detection and texture discrimination (since in both cases curvature either large and/or exhibits a large variation). Results

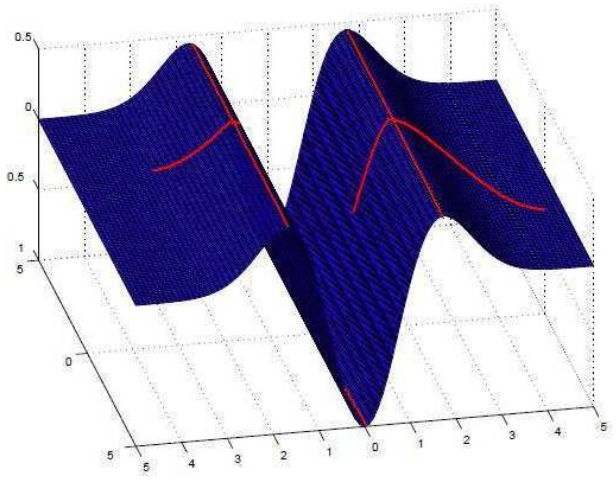


Figure 8: Lines of curvatures on a ridgelet (after [9]).

should be validated using previous work of Brox & Weickert [3] and Lindenberg [17]. It's extension to ridgelets (and curvelets) should be compared with such benchmark works as [6]. Moreover, in view of such works as [4], [15], [16] (to cite only a few), further applications to image compression also impose themselves as naturally stemming from our curvature analysis. In addition, feature extraction is also a natural application for our method, since it allows for a better correlation between the internal scale of he image (i.e. curvature) and wavelets' scale. (In fact, experiments in this direction are currently in progress.)

On the theoretical end of the spectrum, one would like to develop a full multi-curvature analysis framework, where images are constructed using basis functions that are curve-related to one another. This is not an impossible task as it seems, since, as we have already mentioned, we have shown in [24] that image sampling and reconstruction based on their curvature is possible. In fact, in the said paper, we have proven that, in the geometric approach, the *radius of curvature* (see [8]) substitutes for the condition of the Nyquist rate, even in the 1-dimensional case. Since (sectional) curvature is defined as $1/(curvature\ radius)$, the relationship between scale and curvature becomes even clearer, in the light of the results presented herein. Therefore, we aim at presenting a *curvature transform*, akin to the *wavelet transform* and to the *scale transform* of [7]. Of course, in the context of curvatures of ridgelets and curvelets one should consider the appropriate types of transforms.

We conclude with a further natural application of metric curvatures, lying at the confluence of theory and practice, namely to the fractals and their use, in conjunction with wavelets or independent of them, to image processing (see, e.g. [11], [13]). While a metric curvature – namely Menger's metric curvature (see [2], [21]) – was already applied in a purely theoretical context to fractal analysis [20], our geometric method allows for a more flexible and coherent approach, that provides a unified treatment of wavelets (including their extensions mentioned above) and fractals.

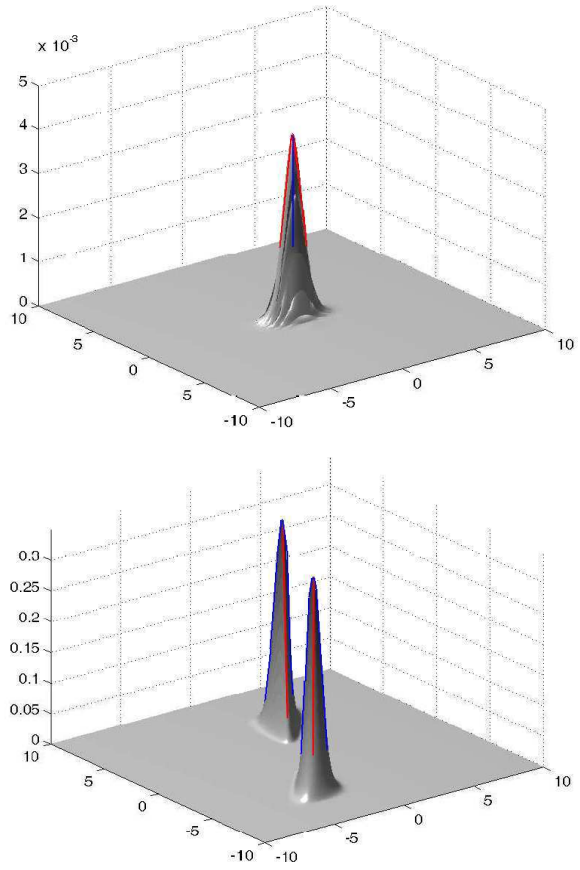


Figure 9: Lines of curvatures on shearlets (after [14]). Note the high positive curvature concentrated at the "apex".

6. Acknowledgments

The authors would like to thank Professor Yehoshua Y. Zeevi for possing the problem, and to Professor Peter Maass, for his constructive critique and encouragement. The first author would also like to thank Professor Shahar Mendelson – his warm support is gratefully acknowledged.

References:

[1] Jean-Pierre Antoine and Laurent Jaques. Measuring a curvature radius with directional wavletes. In J-P. Gazeau, R. Kerner, J-P. Antoine, S. Metens, J-Y. Thibon, editors, *GROUP 24: Physical and Mathematical Aspects of Symmetries, Inst. Phys. Conf. Series 173*, pages 899–904, 2003.

[2] Leonard M. Blumenthal and Karl Menger. *Studies in Geometry* Freeman & co., San Francisco, 1970.

[3] Thomas Brox and Joackim Weickert. A TV flow based local scale estimate and its application to texture discrimination. *Journal of Visual Communication and Image Representation*, 17(5): 1053–1073, October 2006.

[4] A. R. Calderbank, Ingrid Daubechies, Wim Sweldens and Boon-Lock Yeo Lossless image compression us-

ing integer to integer wavelet transforms. In *Proceedings of ICIP 1997*, vol.1, pages 596–599, 1997.

[5] Emmanuel J. Candès and David L. Donoho. Ridgelets: a key to higher-dimensional intermittency? Phil. Trans. R. Soc. Lond. A., 357, 24952509. In L. L. Schumaker et al. editors, *Curves and Surfaces.*, 1999.

[6] Emmanuel J. Candès and David L. Donoho. Curvelets - a surprisingly effective nonadaptive representation for objects with edges. In L. L. Schumaker et al. editors, *Curves and Surfaces.*, pages 1–10, 1999.

[7] Leon Cohen. The Scale Representation *IEEE Trans. Signal Processing*, 41(12): 3275–3292, December 1993.

[8] Manfredo P. do Carmo. *Differential Geometry of Curves and Surfaces*, Prentice-Hall, Englewood Cliffs, N.J., 1976.

[9] David L. Donoho. Ridgelets and Ridge Functions *NSF-SIAM Conference Board in the Mathematical Sciences Lectures*, 2000.

[10] Luc Florack, Bart. M. ter Haar Romeny, Jan J. Koenderink and Max A. Viergever. *Scale and the differential structure of images*, Image Vision Comput. 10(6), 376–388, 1992.

[11] Éric Guérin, Éric Tosan and Atilla Baskurt. Fractal approximation and compression using projected IFS In *Interdisciplinary Approaches in Fractal Analysis, IAFA'2003*, pages 39-45, 2003.

[12] Johannes Haantjes. Distance geometry. Curvature in abstract metric spaces. *Indagationes Math.*, 9: 302-314, 1947

[13] Houssam Hnaidi, Éric Guérin and Samir Akkouche. Fractal/Wavelet representation of objects In *3rd International Conference onInformation and Communication Technologies: From Theory to Applications, ICTTA 2008*, pages 1-5, 2008.

[14] Gitta Kutyniok and Tomas Sauer. From Wavelets to Shearlets and back again In M.Neamtu, L. L. Schumaker, editors, *Approximation Theory XII: San Antonio 2007*, pages 201–209, 2008.

[15] Erwan Le Pennec and Stéphane Mallat. Image compression with geometrical wavelets In *Proceedings of ICIP 2000*, vol.1, pages 661–664, 2000.

[16] Adrian S. Lewis and G. Knowles. Image Compression Using the 2-D Wavelet Transform. *IEEE Transactions on Image Processing* 1(2): 244–250, 1992.

[17] Tony Lindeberg. Edge Detection and Ridge Detection with Automatic Scale Selection. *International Journal of Computer Vision* 30(2): 117–154, 1998.

[18] John M. Lounsbery, Anthony D. DeRose, and Joe Warren. Multiresolution Analysis For Surfaces Of Arbitrary Topological Type *ACM Transactions on Graphics*, 16(1): 34–73 1997.

[19] Yves Meyer. *Wavelets : Algorithms & Applications*. SIAM, University of Michigan, 1993.

[20] Hervé Pajot. *Analytic Capacity, Rectificabilility, Menger Curvature and the Cauchy Integral*. Lecture Notes in Mathematics 1799, Springer-Verlag, Berlin, 2002.

[21] Emil Saucan. Curvature – Smooth, Piecewise-Linear and Metric. In G. Sica, editor, *What is Geometry?, Advanced Studies in Mathematics and Logic*, pages 237–268, 2006.

[22] Emil Saucan. Surface triangulation - the metric approach. Preprint (arxiv:cs.GR/0401023), 2004.

[23] Emil Saucan, and Eli Appleboim. Curvature Based Clustering for DNA Microarray Data Analysis. *Lecture Notes in Computer Science*, 3523:405–412, 2005.

[24] Emil Saucan, Eli Appleboim, and Yehoshua Y Zeevi. Sampling and Reconstruction of Surfaces and Higher Dimensional Manifolds. *Journal of Mathematical Imaging and Vision* 30(1):105–123, 2008.

[25] Emil Saucan, Eli Appleboim, and Yehoshua Y Zeevi. Geometric Approach to Sampling and Communication. Technion CCIT Report #707, November 2008.

[26] Sébastien Valette and Rémy Prost. Wavelet-Based Multiresolution Analysis Of Irregular Surface Meshes. *IEEE Transaction on Visualization and Computer Graphics*, (10)2:113–122, 2004.

[27] Jan-Olov Strömberg. A modified Franklin system and high order spline systems on $\mathbb{R}^n$ as unconditional bases for Hardy spaces. In W. Beckner, editor, *Conference on Harmonic Analysis in honor of A. Zygmund*, pages 475–494, Wadeworth International Group, Belmont, California, 1983.

Analysis of Singularity Lines by Transforms with Parabolic Scaling

Panuvuth Lakhonchai [(1)], Jouni Sampo [(2)] and Songkiat Sumetkijakan [(1)]

(1) Department of Mathematics, Chulalongkorn University, Phyathai Road, Patumwan, Bangkok 10330, Thailand.
(2) Department of Applied Mathematics, Lappeenranta University of Technology, Lappeenranta, Finland.
panuvuth@hotmail.com, jouni.sampo@lut.fi, songkiat.s@chula.ac.th

Abstract:

Using Hart Smith's, curvelet, and shearlet transforms, we investigate L^2 functions with sufficiently smooth background and present here sufficient and necessary conditions, which include the special case with 1-dimensional singularity line. Specifically, we consider the situation where regularity on a line in a non-parallel direction is much lower than directional regularity along the line in a neighborhood and how this is reflected in the behavior of the three transforms.

1. Introduction

Wavelet transforms, both continuous and discrete, have proved to be a very efficient tool in detecting point singularities. However, due to its isotropic scaling, wavelet transforms are not ideal tools in detecting one-dimensional singularities like singularity lines or curves. Recently, wavelet-like transforms with parabolic scaling, such as Hart Smith's and curvelet transforms, were introduced and applied successfully in edge detection. Our goal is then to investigate how these transforms can be used in detecting point, line, and curve singularities. New necessary and new sufficient conditions for an $L^2(\mathbb{R}^2)$ function to possess Hölder regularity, uniform and pointwise, with exponent $\alpha > 0$ are given. Similar to the characterization of Hölder regularity by the continuous wavelet transform, the conditions here are in terms of bounds of the Smith and curvelet transforms across fine scales. However, due to the parabolic scaling, the sufficient and necessary conditions differ in both the uniform and pointwise cases, with larger gap in pointwise regularities. Naturally, global conditions for pointwise singularities can be weakened. We then investigate functions with sufficiently smooth background in one direction and potential singularity in the perpendicular (non-parallel) direction. Specifically, sufficient and necessary conditions, which include the special case with one-dimensional singularity line, are derived for pointwise Hölder exponent. Inside their "cones" of influence, these conditions are practically the same, giving near-characterization of direction of singularity.

2. Directional Regularity

We shall restrict our definition to a real-valued function f of two variables. Generalization to a function of several variables is straightforward. For a given positive exponent α not in $\mathbb{N}$, its pointwise, uniform, and directional Hölder (or Lipschitz) regularities are defined as follows. Fix a point $\boldsymbol{u} \in \mathbb{R}^2$ at which regularity is under investigation. f is said to be *pointwise Hölder regular with exponent α at $\boldsymbol{u}$*, denoted by $f \in C^\alpha(\boldsymbol{u})$, if there exists a polynomial $P_{\boldsymbol{u}}$ of degree less than α and a constant $C = C_{\boldsymbol{u}}$ such that for all $\boldsymbol{x}$ in a neighborhood of $\boldsymbol{u}$

$$|f(\boldsymbol{x}) - P_{\boldsymbol{u}}(\boldsymbol{x} - \boldsymbol{u})| \le C\|\boldsymbol{x} - \boldsymbol{u}\|^\alpha. \tag{1}$$

If there exists a uniform constant C so that for all $\boldsymbol{u}$ in an open subset Ω of $\mathbb{R}^2$ there is a polynomial $P_{\boldsymbol{u}}$ of degree less than α such that (1) holds for all $\boldsymbol{x} \in \Omega$, then we say that f is *uniformly Hölder regular with exponent α on Ω* or $f \in C^\alpha(\Omega)$. The *uniform Hölder exponent* of f on Ω is defined to be

$$\alpha_l(\Omega) := \sup\{\alpha \colon f \in C^\alpha(\Omega)\}, \tag{2}$$

and the *pointwise Hölder exponent* is defined in an analogous manner. Following [9], the *local Hölder exponent* of f at $\boldsymbol{u}$ is defines as

$$\alpha_l(\boldsymbol{u}) = \lim_{n\to\infty} \alpha_l(I_n).$$

where $\{I_n\}_{n\in\mathbb{N}}$ is a family of nested open sets in $\mathbb{R}^2$, i.e. $I_{n+1} \subset I_n$, with intersection $\cap_n I_n = \{\boldsymbol{u}\}$.
In order to define directional regularity, let $\boldsymbol{v} \in \mathbb{R}^d$ be a fixed unit vector representing a direction and $\boldsymbol{u}$ be a point in $\mathbb{R}^d$. f is said to be *pointwise Hölder regular with exponent α at $\boldsymbol{u}$ in the direction $\boldsymbol{v}$*, denoted by $f \in C^\alpha(\boldsymbol{u}; \boldsymbol{v})$, if there exist a constant $C = C_{\boldsymbol{u},\boldsymbol{v}}$ and a polynomial $P_{\boldsymbol{u},\boldsymbol{v}}$ of degree less than α such that

$$|f(\boldsymbol{u} + \lambda\boldsymbol{v}) - P_{\boldsymbol{u},\boldsymbol{v}}(\lambda)| \le C|\lambda|^\alpha \tag{3}$$

holds for all λ in a neighborhood of $0 \in \mathbb{R}$. We next define directional regularity on a set $\Omega_1 \subseteq \mathbb{R}^2$. Let Ω_2 be an open neighborhood of Ω_1 representing a set on which the Hölder estimate holds. Then f is said to be in $C^\alpha(\Omega_1, \Omega_2; \boldsymbol{v})$ if there exists a constant $C = C_{\boldsymbol{v}}$ so that for all $\boldsymbol{u} \in \Omega_1$ there is a polynomial $P_{\boldsymbol{u},\boldsymbol{v}}$ of degree less than α such that (3) holds for all $\lambda \in \mathbb{R}$ with $\boldsymbol{u} + \lambda\boldsymbol{v} \in \Omega_2$. If $\Omega_1 = \Omega_2$, then we denote $C^\alpha(\Omega_1, \Omega_2; \boldsymbol{v})$ simply by $C^\alpha(\Omega_1; \boldsymbol{v})$. Of course, the *directional pointwise and uniform Hölder exponents* could be defined in the same way as (2). In the pointwise case, this directional

Hölder exponent measures one-dimensional regularity of f at $\boldsymbol{u}$ on the line passing through $\boldsymbol{u}$ and parallel with $\boldsymbol{v}$. See [5]. For $C^\alpha(\Omega_1, \Omega_2; \boldsymbol{v})$, the set Ω_1 in our context of line singularity will usually be a line and $\boldsymbol{v}$ points in a direction that is nonparallel with the line. In this situation, $f \in C^\alpha(\Omega_1, \Omega_2; \boldsymbol{v})$ has a ridge along the line provided that hte regularity in the direction of the line is sufficiently high. See Theorem 4.

3. Three Transforms with Parabolic Scaling

3.1 Hart Smith Transform

Originally defined in [10], the Hart Smith transform was described in [1, 2] as follows. For a given $\varphi \in L^2(\mathbb{R}^2)$, we define

$$\varphi_{a\boldsymbol{b}\theta}(\boldsymbol{x}) = a^{-\frac{3}{4}} \varphi\left(\mathrm{D}_{\frac{1}{a}} \mathrm{R}_{-\theta}\left(\boldsymbol{x} - \boldsymbol{b}\right)\right),$$

for $\theta \in [0, 2\pi)$, $\boldsymbol{b} \in \mathbb{R}^2$, and $0 < a < a_0$, where a_0 is a fixed coarsest scale, $\mathrm{D}_{\frac{1}{a}} = \mathrm{diag}\left(\frac{1}{a}, \frac{1}{\sqrt{a}}\right)$, and $R_{-\theta}$ is the matrix affecting planar rotation of θ radians in clockwise direction. Hart Smith transform can then be defined as

$$\overline{\Gamma}_f\left(a, \boldsymbol{b}, \theta\right) := \left\langle \varphi_{a\boldsymbol{b}\theta}, f \right\rangle.$$

This gives a true affine transform that uses parabolic scaling. For each scale a and direction θ, let us define the norm

$$\|\boldsymbol{v}\|_{a,\theta} := \left\| D_{\frac{1}{a}} R_{-\theta} \boldsymbol{v} \right\| \quad \text{for } \boldsymbol{v} \in \mathbb{R}^2.$$

We define vector $\boldsymbol{v}_\theta := R_\theta (0,1)^T$ so that $\boldsymbol{v}_\theta$ is parallel to the major axis of the ellipse $\|\boldsymbol{v}\|_{a,\theta} = 1$.

Reconstruction Formula [10, 1, 2]
There exists a Fourier multiplier M of order 0 so that whenever $f \in L^2(\mathbb{R}^2)$ is a high-frequency function supported in frequency space $\|\boldsymbol{\xi}\| > \frac{2}{a_0}$, then, in $L^2(\mathbb{R}^2)$

$$\begin{aligned} f &= \int_0^{a_0} \int_0^{2\pi} \int_{\mathbb{R}^2} \left\langle \varphi_{a\boldsymbol{b}\theta}, Mf \right\rangle \varphi_{a\boldsymbol{b}\theta} \, d\boldsymbol{b} \, d\theta \, \frac{da}{a^3} \quad (4) \\ &= \int_0^{a_0} \int_0^{2\pi} \int_{\mathbb{R}^2} \left\langle \varphi_{a\boldsymbol{b}\theta}, f \right\rangle M\varphi_{a\boldsymbol{b}\theta} \, d\boldsymbol{b} \, d\theta \, \frac{da}{a^3}. \end{aligned}$$

3.2 Continuous Curvelet Transform

Following Candès and Donoho[1, 2], the continuous curvelet transform (CCT) is defined in the polar coordinates (r, ω) of the Fourier domain. Let W be a positive real-valued C^∞ function supported inside $\left(\frac{1}{2}, 2\right)$, called a *radial window*, and let V be a real-valued C^∞ function supported on $[-1, 1]$, called an *angular window*, for which the following admissibility conditions hold:

$$\int_0^\infty W(r)^2 \frac{dr}{r} = 1 \quad \text{and} \quad \int_{-1}^1 V(\omega)^2 \, d\omega = 1. \quad (5)$$

At each scale a, $0 < a < a_0$, $\gamma_{a\boldsymbol{0}0}$ is defined by

$$\widehat{\gamma_{a\boldsymbol{0}0}}\left(r\cos(\omega), r\sin(\omega)\right) = a^{\frac{3}{4}} W\left(ar\right) V\left(\omega/\sqrt{a}\right)$$

for $r \geq 0$ and $\omega \in [0, 2\pi)$. For each $0 < a < a_0$, $\boldsymbol{b} \in \mathbb{R}^2$, and $\theta \in [0, 2\pi)$, a *curvelet* $\gamma_{a\boldsymbol{b}\theta}$ is defined by

$$\gamma_{a\boldsymbol{b}\theta}(\boldsymbol{x}) = \gamma_{a\boldsymbol{0}0}\left(R_\theta\left(\boldsymbol{x} - \boldsymbol{b}\right)\right), \quad \text{for } \boldsymbol{x} \in \mathbb{R}^2. \quad (6)$$

The continuous curvelet transform of $f \in L^(\mathbb{R}^2)$ is

$$\Gamma_f\left(a, \boldsymbol{b}, \theta\right) = \left\langle \gamma_{a\boldsymbol{b}\theta}, f \right\rangle$$

for $0 < a < a_0$, $\boldsymbol{b} \in \mathbb{R}^2$, and $\theta \in [0, 2\pi)$.
The admissibility conditions (5) and the polar coordinate design of curvelets yield the following:

Reconstruction formula [2]
There exists a bandlimited purely radial function Φ such that for all $f \in L^2(\mathbb{R}^2)$,

$$f = \tilde{f} + \int_0^{a_0} \int_0^{2\pi} \int_{\mathbb{R}^2} \left\langle \gamma_{a\boldsymbol{b}\theta}, f \right\rangle \gamma_{a\boldsymbol{b}\theta} \, d\boldsymbol{b} \, d\theta \, \frac{da}{a^3}, \quad (7)$$

where $\tilde{f} = \int_{\mathbb{R}^2} \left\langle \Phi_{\boldsymbol{b}}, f \right\rangle \Phi_{\boldsymbol{b}} \, d\boldsymbol{b}$ and $\Phi_{\boldsymbol{b}}(\boldsymbol{x}) = \Phi(\boldsymbol{x} - \boldsymbol{b})$.
For analysis of singularities of f, the low frequency part $\tilde{f}$ is not an issue as it is always C^∞. Unlike Smith transform, curvelet transform does not use a true affine parabolic scaling as a slightly different generating function $\gamma_{a\boldsymbol{0}0}$ is used at each scale $a > 0$.

3.3 Continuous Shearlet Transform

We will follow mainly the definitions and notations in G. Kutyniok and D. Labate[6]. Let $\psi_1, \psi_2 \in L^2(\mathbb{R})$ and $\psi \in L^2(\mathbb{R}^2)$ be given by

$$\hat{\psi}(\xi_1, \xi_2) = \hat{\psi}_1(\xi_1)\hat{\psi}_2\left(\frac{\xi_2}{\xi_1}\right), \quad \xi_1 \neq 0, \xi_2 \in \mathbb{R}, \quad (8)$$

where ψ_1 satisfies the admissibility condition and $\hat{\psi}_1 \in C_0^\infty(\mathbb{R})$ with $\mathrm{supp}\,\hat{\psi}_1 \subset [-2, -\frac{1}{2}] \cup [\frac{1}{2}, 2]$ while $\hat{\psi}_2 \in C_0^\infty(\mathbb{R})$ with $\mathrm{supp}\,\hat{\psi}_2 \subset [-1, 1]$, $\hat{\psi}_2 > 0$ on $(-1, 1)$, and $\|\psi\|_2 = 1$. Given such a *shearlet function* ψ, a *continuous shearlet system* is the family of functions $\psi_{as\boldsymbol{t}}$, $a \in \mathbb{R}^+$, $s \in \mathbb{R}$, $\boldsymbol{t} \in \mathbb{R}^2$, where

$$\psi_{as\boldsymbol{t}} = a^{-\frac{3}{4}} \psi\left(\mathrm{D}_a^{-1} \mathrm{B}_s^{-1}(\cdot - \boldsymbol{t})\right)$$

where B_s is the *shear matrix* $\begin{pmatrix} 1 & -s \\ 0 & 1 \end{pmatrix}$ and D_a is the diagonal matrix $\begin{pmatrix} a & 0 \\ 0 & \sqrt{a} \end{pmatrix}$. The *continuous shearlet transform* of f is then defined for such $(a, s, \boldsymbol{t})$ by

$$\mathcal{SH}_\psi f(a, s, \boldsymbol{t}) = \left\langle f, \psi_{as\boldsymbol{t}} \right\rangle.$$

Many properties of the continuous shearlet are more evident in the frequency domain. So we note here that each $\hat{\psi}_{as\boldsymbol{t}}$ is supported on the set

$$\left\{(\xi_1, \xi_2) \colon \frac{1}{2a} \leq |\xi_1| \leq \frac{2}{a}, \left|\frac{\xi_2}{\xi_1} - s\right| \leq \sqrt{a}\right\}.$$

Reconstruction Formula [6]
Let $\psi \in L^2(\mathbb{R}^2)$ be a shearlet function. Then, for all $f \in L^2(\mathbb{R}^2)$,

$$f = \int_{\mathbb{R}^2} \int_{\mathbb{R}} \int_{\mathbb{R}^+} \left\langle \psi_{as\boldsymbol{t}}, f \right\rangle \psi_{as\boldsymbol{t}} \, \frac{da}{a^3} \, ds \, d\boldsymbol{t} \text{ in } L^2. \quad (9)$$

If $\operatorname{supp}\hat{f} \subset C = \left\{(\xi_1,\xi_2)\colon |\xi_1| \geq 2 \text{ and } \left|\frac{\xi_2}{\xi_1}\right| \leq 1\right\}$, then

$$f = \int_{\mathbb{R}^2}\int_{-2}^{2}\int_{0}^{1} \langle \psi_{ast}, f\rangle\, \psi_{ast}\, \frac{da}{a^3}\, ds\, d\boldsymbol{t} \text{ in } L^2. \quad (10)$$

Even though the second reconstruction formula (10) is valid only for functions with frequency support in the union C of two infinite horizontal trapezoids, it has the advantage that the integral involves only scales a and shear parameters s in bounded sets. A complementary shearlet system $\psi^{(v)}_{ast}$ can be similarly defined so that one has a reconstruction formula which is valid for f with $\operatorname{supp}\hat{f} \subset C^{(v)} = \left\{(\xi_1,\xi_2)\colon |\xi_2| \geq 2 \text{ and } \left|\frac{\xi_2}{\xi_1}\right| > 1\right\}$. Finally, every $f \in L^2(\mathbb{R}^2)$ can be decomposed into three functions with frequency supports in C, $C^{(v)}$, and $W = [-2,2]^2$. The former two functions can then be reconstructed from ψ_{ast} and $\psi^{(v)}_{ast}$ respectively, while the latter is C^∞. Therefore, regularity analysis can be carried out by considering the continuous shearlet transform with respect to these two shearlet systems. For more details, see [6].

4. Common Properties of the Transforms

We shall suppose from this point onward that $\hat{\varphi} \in C^\infty$ and that there exist $C_1' > C_1' > 0$ and $C_2 > 0$ such that $\operatorname{supp}(\hat{\varphi}) \subset ([-C_1', -C_1] \cup [C_1, C_1']) \times [-C_2, C_2]$. This assumption ensures that all our three kernel functions, Hart Smith, curvelet, and shearlet functions, have Fourier supports away from the Y-axis, which in turns results in crucial properties needed to prove our main results.

4.1 Vanishing Directional Moments

A function f of two variables is said to have an *L-order vanishing directional moments along a direction* $\boldsymbol{v} = (v_1, v_2)^T \neq \mathbf{0}$ if

$$\int_{\mathbb{R}} b^n f(b\boldsymbol{v}+\boldsymbol{w})db = 0, \quad \text{for all } \boldsymbol{w} \in \mathbb{R}^2 \text{ and } 0 \leq n < L.$$

Lemma 1: Let $\boldsymbol{v} = (v_1, v_2)^T$ be a unit vector.

1. There exists $C < \infty$ (independent of a, $\boldsymbol{b}$ and θ) such that if $|\theta + \arctan(\frac{v_1}{v_2})| \geq C\sqrt{a}$ then the curvelet functions $\gamma_{ab\theta}$ and the Smith functions $\varphi_{ab\theta}$ and $M\varphi_{ab\theta}$ have vanishing directional moments of any order $L < \infty$ along the direction $\boldsymbol{v}$.

2. If $\left|s + \frac{v_1}{v_2}\right| > \sqrt{a}$ then the shearlet functions ψ_{ast} have vanishing directional moments of any order $L < \infty$ along the direction $\boldsymbol{v}$. Here, if v_2 is 0 then $\frac{v_1}{v_2}$ are treated as ∞ so that the assumed inequality holds for all $a \in (0,1)$ and $s \in [-2,2]$, hence ψ_{ast} has vanishing directional moments of any order $L < \infty$ along the direction $\boldsymbol{v} = (v_1, 0)$.

4.2 Smoothness and Decay Properties

Lemma 2: For each $N = 1, 2, \ldots$ there is a constant C_N such that for all $\boldsymbol{x} \in \mathbb{R}^2$ and $\nu \in \mathbb{N}_0^2$

$$|\partial^\nu \gamma_{ab\theta}(\boldsymbol{x})| \leq \frac{C_N a^{-3/4-|\nu|}}{1 + \left\|D_{\frac{1}{a}} R_{-\theta}(\boldsymbol{x}-\boldsymbol{b})\right\|^{2N}} \quad (11)$$

and

$$|\partial^\nu \psi_{ast}(\boldsymbol{x})| \leq \frac{C_N a^{-3/4-|\nu|}(\sqrt{a}+|s|)^{\nu_2}}{1 + \left\|\mathrm{D}_{1/a}\mathrm{B}_{-s}(\boldsymbol{x}-\boldsymbol{t})\right\|^{2N}}. \quad (12)$$

Moreover, (11) also holds for functions $\varphi_{ab\theta}$ and $M\varphi_{ab\theta}$.

5. Singularity Lines

Let $\phi_{ab\theta}$ denote any of the $\gamma_{ab\theta}$, $\varphi_{ab\theta}$, or $M\varphi_{ab\theta}$. Let us quote the following results.[8, 7]

Theorem 1: Let $f \in L^2(\mathbb{R}^2)$, $\boldsymbol{u} \in \mathbb{R}^2$, and assume that $\alpha > 0$ is not an integer. If there exist $\alpha' < 2\alpha$, $\theta_0 \in [0, 2\pi]$, and $A, C < \infty$ such that $|\langle \phi_{ab\theta}, f\rangle|$ is bounded by

$$\begin{cases} Ca^{\alpha+\frac{5}{4}}\left(1 + \left\|\dfrac{\boldsymbol{b}-\boldsymbol{u}}{a^{1/2}}\right\|^{\alpha'}\right), & \text{if } |\theta-\theta_0| \geq A\sqrt{a} \\ Ca^{\alpha+\frac{3}{4}}\left(1 + \left\|\dfrac{\boldsymbol{b}-\boldsymbol{u}}{a^{1/2}}\right\|^{\alpha'}\right), & \text{if } |\theta-\theta_0| \leq A\sqrt{a} \end{cases}$$

for all $a \in (0, a_0)$, $\boldsymbol{b} \in \mathbb{R}^2$, and $\theta \in [0, 2\pi)$, then $f \in C^\alpha(\boldsymbol{u})$.

Theorem 2: Let $f \in L^2(\mathbb{R}^2)$, $\boldsymbol{u} \in \mathbb{R}^2$, and assume that $\alpha > 0$ is not an integer. If there exist $\alpha' < 2\alpha$, $-2 \leq s_0 \leq 2$, and $C, C' < \infty$ such that, for each $0 < a < 1$, $-2 \leq s \leq 2$, and $\boldsymbol{t} \in \mathbb{R}^2$, $|\langle \psi_{ast}, P_{C_1} f\rangle|$ is bounded by

$$\begin{cases} Ca^{\alpha+\frac{5}{4}}\left(1 + \left\|\dfrac{\boldsymbol{t}-\boldsymbol{u}}{a^{1/2}}\right\|^{\alpha'}\right), & \text{if } |s-s_0| > C'\sqrt{a}, \\ Ca^{\alpha+\frac{3}{4}}\left(1 + \left\|\dfrac{\boldsymbol{t}-\boldsymbol{u}}{a^{1/2}}\right\|^{\alpha'}\right), & \text{if } |s-s_0| \leq C'\sqrt{a}, \end{cases} \quad (13)$$

and

$$\left|\left\langle \psi^{(v)}_{ast}, P_{C_2} f\right\rangle\right| \leq Ca^{\alpha+\frac{5}{4}}\left(1 + \left\|\frac{\boldsymbol{t}-\boldsymbol{u}}{a^{1/2}}\right\|^{\alpha'}\right), \quad (14)$$

then $f \in C^\alpha(\boldsymbol{u})$. Similar statement holds if the inequality (13) holds for $\left\langle \psi^{(v)}_{ast}, P_{C_2} f\right\rangle$ and the inequality (14) holds for $\langle \psi_{ast}, P_{C_1} f\rangle$.

Theorem 3 Let f be bounded with local Hölder exponent $\alpha \in (0, 1]$ at point $\boldsymbol{u}$ and $f \in C^{2\alpha+1+\varepsilon}(\mathbb{R}^2, \boldsymbol{v}_{\theta_0})$ for some $\theta_0 \in [0, 2\pi)$ with any fixed $\varepsilon > 0$. Then there exist $\alpha' \in [\alpha-\varepsilon, \alpha]$ and $A, C < \infty$ such that for $a > 0$ and $\boldsymbol{b} \in \mathbb{R}^2$, $|\langle \phi_{ab\theta}, f\rangle|$ is bounded by

$$\begin{cases} Ca^{\alpha+\frac{5}{4}}, & \text{if } |\theta-\theta_0| \geq A\sqrt{a}, \\ Ca^{\alpha'+\frac{3}{4}}\left(1 + \left\|\dfrac{\boldsymbol{b}-\boldsymbol{u}}{a}\right\|^{\alpha'}\right), & \text{if } |\theta-\theta_0| \leq A\sqrt{a}. \end{cases}$$

For $s_0 \in [-2,2]$ and $\boldsymbol{u} = (u_1, u_2) \in \mathbb{R}^2$, let $\Gamma_{\boldsymbol{u}}$ denote the vertical line passing though $\boldsymbol{u}$ and $\Gamma_{\boldsymbol{u},s_0}$ denote the line passing through $\boldsymbol{u}$ with slope $-\frac{1}{s_0}$. Observe that we may write $\Gamma_{\boldsymbol{u}} = \Gamma_{\boldsymbol{u},0}$ so that $(x_1, x_2) \in \Gamma_{\boldsymbol{u},s_0}$ if and only if $x_1 = -s_0(x_2 - u_2) + u_1$. Recall that if $\Gamma \subseteq \mathbb{R}^2$ and $\rho > 0$, then $\Gamma(\rho)$ is the ρ-neighborhood of Γ, i.e. the set of all points whose distance to Γ is less than ρ.

Theorem 4 Let $f \in C^{\alpha}(\Gamma_{\boldsymbol{u},s_0}, \Gamma_{\boldsymbol{u},s_0}(\rho); (1,0))$ and bounded for some $\alpha \in (0,1]$, $\boldsymbol{u} \in \mathbb{R}^2$, $s_0 \in [-2,2]$ and $\rho > 1$. Suppose also that f is in $C^{2\alpha+1+\varepsilon}(\Gamma_{\boldsymbol{u},s_0}(\rho); \mathrm{B}_{s_0}(0,1))$ for some fixed $\varepsilon > 0$. Then there exists $C < \infty$ such that if $0 < a < a_0 < 1$ and $\boldsymbol{t} \in \Gamma_{\boldsymbol{u}}(r)$ with $r < \rho/2$ and $s \in [-2,2]$, the continuous shearlet transform $\langle \psi_{as\boldsymbol{t}}, f \rangle$ is bounded in magnitude by

$$\begin{cases} Ca^{\alpha+\frac{5}{4}}, & \text{if } |s - s_0| > \sqrt{a}, \\ Ca^{\alpha+\frac{3}{4}} \left(1 + \left|\frac{d_{s_0}(\boldsymbol{t},\boldsymbol{u})}{a}\right|^{\alpha}\right), & \text{if } |s - s_0| \leq \sqrt{a}, \end{cases}$$

where $d_{s_0}(\boldsymbol{t},\boldsymbol{u}) = |t_1 + s_0 t_2 - u_1 - s_0 u_2|$ denotes the distance between the parallel lines with slope $-\frac{1}{s_0}$ (vertical line if $s_0 = 0$) and passing through $\boldsymbol{t}$ and $\boldsymbol{u}$ respectively.

Edge analysis has been done successfully using the continuous shearlet transform ([11, 4, 3, 6]). They consider the shearlet transform of the characteristic function of a set with piecewise smooth boundary and found that, at a regular boundary point $\boldsymbol{t}$, the shearlet transform decays like $a^{3/4}$ if $s = s_0 = \pm\frac{v_1}{v_2}$ and decays rapidly at other $s \neq s_0$, where $\boldsymbol{v} = (v_1, v_2)$ is the normal vector of the boundary curve at $\boldsymbol{t}$. Since this characteristic function has Hölder exponent 0 (bounded and discontinuous) at any boundary point in the normal direction, this decay rate of $a^{3/4}$ at $s = s_0 = 0$ agrees with that of Theorem 4. However, when $s_0 \neq 0$ the two directions in Theorem 4 along which regularity is assumed are not perpendicular. More comparisons of our results and the aforementioned work are needed.

References:

[1] Emmanuel J. Candès and David L. Donoho. Continuous curvelet transform. I: Resolution of the wavefront set. *Appl. Comput. Harmon. Anal.*, 19(2):162–197, 2005.

[2] Emmanuel J. Candès and David L. Donoho. Continuous curvelet transform. II: Discretization and frames. *Appl. Comput. Harmon. Anal.*, 19(2):198–222, 2005.

[3] K. Guo, Labate D., and W-Q. Lim. Edge analysis and identification using the continuous shearlet transform. *Appl. Comput. Harmon. Anal.*, 2008. In Press.

[4] K. Guo and D. Labate. Characterization and analysis of edges using the continuous shearlet transform. 2008. Preprint.

[5] S. Jaffard. Multifractal functions: Recent advances and open problems. *Menuscript*, 2004.

[6] G. Kutyniok and D. Labate. Resolution of the wavefront set using continuous shearlets. *Trans. AMS.*, 105(1):157–175, 2007.

[7] P. Lakhonchai, J. Sampo, and S. Sumetkijakan. Shearlet transforms and hölder regularities. 2009. Preprint.

[8] J. Sampo and S. Sumetkijakan. Estimations of Hölder regularities and direction of singularity by Hart Smith and curvelet transforms. *Journal of Fourier Analysis and Applications*, 15(1):58–79, 2009.

[9] S. Seuret and J. Lévy Véhel. The local Hölder function of a continuous function. *Appl. Comput. Harmon. Anal.*, 13(3):263–276, 2002.

[10] Hart F. Smith. A Hardy space for Fourier integral operators. *J. Geom. Anal.*, 8(4):629–653, 1998.

[11] S. Yi, D. Labate, G.R. Easley, and H. Krim. Edge detection and processing using shearlets. 2008. Preprint.

Geometric Separation using a Wavelet-Shearlet Dictionary

David L. Donoho [1] and Gitta Kutyniok [2]

(1) Department of Statistics, Stanford University, Stanford, CA 94305, USA.
(2) Institute of Mathematics, University of Osnabrück, 49069 Osnabrück, Germany.
donoho@stanford.edu, kutyniok@uni-osnabrueck.de

Abstract:

Astronomical images of galaxies can be modeled as a superposition of pointlike and curvelike structures. Astronomers typically face the problem of extracting those components as accurate as possible. Although this problem seems unsolvable – as there are two unknowns for every datum – suggestive empirical results have been achieved by employing a dictionary consisting of wavelets and curvelets combined with ℓ_1 minimization techniques. In this paper we present a theoretical analysis in a model problem showing that accurate geometric separation can be achieved by ℓ_1 minimization. We introduce the notions of *cluster coherence* and clustered sparse objects as a machinery to show that the underdetermined system of equations can be stably solved by ℓ_1 minimization. We prove that not only a radial wavelet-curvelet dictionary achieves nearly-perfect separation at all sufficiently fine scales, but, in particular, also an orthonormal wavelet-shearlet dictionary, thereby proposing this dictionary as an interesting alternative for geometric separation of pointlike and curvelike structures. To derive this final result we show that curvelets and shearlets are sparsity equivalent in the sense of a finite p-norm ($0 < p \leq 1$) of the cross-Grammian matrix.

1. Introduction

Cosmological data analysts face tasks of *geometric separation*. Gravitation, acting over time, drives an initially quasi-uniform distribution of matter in 3D to concentrate near lower-dimensional structures: points, filaments, and sheets. It would be desirable to process single 'maps' of matter density and somehow extract three 'pure' maps containing just the points, just the filaments, and just the sheets around which matter is concentrating. However, this problem contains three unknowns for every datum which seems impossible to solve on mathematical grounds.

Surprisingly, astronomer Jean-Luc Starck and collaborators have recently been empirically successful in numerical experiments with component separation. They used two or more overcomplete frames, each one specially adapted to particular geometric structures, and were able to obtain separation despite the fact that the underlying system of equations is highly underdetermined.

Here we analyze such approaches in a mathematical framework where we can show that success stems from an interplay between geometric properties of the objects to be separated, and the harmonic analysis for singularities of various geometric types.

1.1 Singularities and Sparsity

As a mathematical idealization of 'image', consider a Schwartz distribution f with domain $\mathbf{R}^2$. The distribution f will be given singularities with specified geometry: points and curves.

We plan to represent such an 'image' using tools of harmonic analysis; in particular bases and frames. While many such representations are conceivable, we are interested here just in those bases or frames which can sparsely represent f.

The type of basis which best sparsifies f depends on the geometry of its singularities. If the singularities occur at a finite number of (variable) points, then *wavelets* give what is, roughly speaking, an optimally sparse representation. If the singularities occur at a finite number of smooth curves, then one of the recently studied directional multiscale representations (*curvelets* or *shearlets*) will do the best job of sparsification.

Since we are concerned with f being a mixture of content types, i.e., points and curves, presumably *both* systems are needed to represent f sparsely.

1.2 Minimum ℓ_1 Decomposition and Perfect Separation

In the early 1990's, R. R. Coifman, Wickerhauser and coworkers became interested in the problem of representing signals using more than one basis and started a first heuristic exploration motivated intuitively, see [5]. A few years later, one of us worked with S .S. Chen to develop a formal, optimization-based approach to the multiple-basis representation problem [4]. Given bases Φ_i, $i = 1, 2$, one solves the following problem

$$(\mathrm{BP}) \quad \min \|\alpha_1\|_1 + \|\alpha_2\|_1 \text{ subject to } S = \Phi_1\alpha_1 + \Phi_2\alpha_2,$$

thereby exploiting that the ℓ_1 norm has a tendency to find sparse solutions when they exist. This can be regarded as the starting point for ℓ_1 decomposition techniques. For *theoretical work* on this topic we refer to, e.g., [6, 10, 15, 16], and for *empirical work* see, for instance, [9,12,14,15].

For further references we would like to mention the survey paper [1].

1.3 A Geometric Separation Problem

The work just cited, while suggestive and inspiring, concerns discretely indexed signal/image processing, and so is either empirical or else rigorously analytical but not directly relevant to *geometric* separation tasks, which will involve always continuum ideas.
In this paper we develop related methods in a mathematical setting where the notion of successful separation can be made definitionally precise and can be established by mathematical analysis. For this, we pose a simple but clear model problem of geometric separation.
Consider a 'pointlike' object $\mathcal{P}$ made of point singularities:

$$\mathcal{P} = \sum_{i=1}^{P} |x - x_i|^{-1}.$$

Consider as well a curvelike object $\mathcal{C}$, a singularity along a closed curve $\tau : [0,1] \mapsto \mathbf{R}^2$:

$$\mathcal{C} = \int \delta_{\tau(t)} dt,$$

where δ_x is the usual Dirac Delta at x. By this choice, we arrange that one of the two distributions does not become dramatically larger than the other as we go to finer and finer scales; rather the ratio of energies is more or less independent of scale. This makes the separation problem challenging at every scale.
Now assume that we observe the 'Signal'

$$f = \mathcal{P} + \mathcal{C}, \tag{1}$$

however, the distributions $\mathcal{P}$ and $\mathcal{C}$ are unknown to us. The *Geometric Separation Problem* now consists in recovering $\mathcal{P}$ and $\mathcal{C}$ from knowledge of f.

1.4 Two Geometric Frames

We focus on two pairs of overcomplete systems for representing the object f:

- *Radial Wavelets* – a tight frame with perfectly isotropic generating elements.
- *Curvelets* – a highly directional tight frame with increasingly anisotropic elements at fine scales.

as well as the pair

- *Orthonormal Separable Meyer Wavelets* – an orthonormal basis of perfectly isotropic generating elements.
- *Shearlets* – a highly directional tight frame with increasingly anisotropic elements at fine scales and a unified treatment of both the continuous and digital setting.

We pick these because, as is well known, point singularities are coherent in wavelets and curvilinear singularities are coherent in curvelets/shearlets. For the precise definitions we refer to [2, 3], [11, 13], as well as [7].

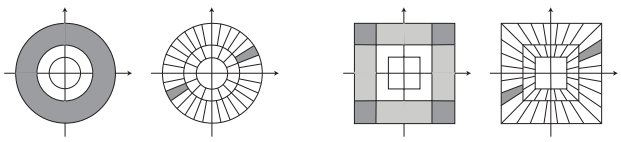


Figure 1: Frequency tilings of radial wavelets and curvelets as well as of orthonormal wavelets and shearlets (from left to right).

Since the scaling subband of each pair are similar as illustrated in Figure 1, we can define two families of filters $(F_j^C)_j$ and $(F_j^S)_j$ which allows to decompose a function f into pieces f_j^C (resp. f_j^S) with different scales j. The piece f_j^C (resp. f_j^S) at subband j arises from filtering f using F_j^C (resp. F_j^S):

$$f_j^C = F_j^C \star f \text{ and } f_j^S = F_j^S \star f,$$

so that the Fourier transform $\hat{f}_i^C$ (resp. $\hat{f}_j^S$) is supported in the scaling subband of scale j of the associated pair of tight frames. The filters are defined in such as way, that we can reconstruct the original function from these pieces using the formula

$$f = \sum_j F_j^C \star f_j^C = \sum_j F_j^S \star f_j^S, \quad f \in L^2(\mathbf{R}^2).$$

For the precise construction of those filters and further properties, we refer to [7].
We can now use these tools to attack the Geometric Separation Problem scale-by-scale. For this, we filter the model problem (1) to derive the sequences of filtered images

$$f_j^C = \mathcal{P}_j^C + \mathcal{C}_j^C \text{ and } f_j^S = \mathcal{P}_j^S + \mathcal{C}_j^S \text{ for all scales } j. \tag{2}$$

1.5 Outline

In Section 2 we will develop and analyze the decomposition technique based on ℓ_1 minimization we intend to employ, first in a very general Hilbert space setting. These results will then be applied to the scale-dependent Geometric Separation Problem (2) proving that the radial wavelet-curvelet as well as the orthonormal wavelet-shearlet dictionary achieves nearly-perfect separation at all sufficient fine scales (Theorems 1 and 3). The sparsity equivalence between curvelets and shearlets we derive in Subsection 3.2 thereby allows transference of this result from the radial wavelet-curvelet to the orthonormal wavelet-shearlet dictionary.

2. General Component Separation

We now first study the behavior of ℓ^1 minimization in the general two-frame case. Suppose we have two tight frames Φ_1, Φ_2 in a Hilbert space $\mathcal{H}$, and a signal vector $S \in \mathcal{H}$. We know *a priori* that there exists a decomposition

$$S = S_1^0 + S_2^0,$$

where S_1^0 is sparse in Φ_1 and S_2^0 is sparsely represented by Φ_2. Our analysis will center on the use of cluster coherence to exploit the geometric structure of the sparse expansions rather than merely the fact that the vector is sparse.

2.1 Cluster Coherence

Typically, separation results employ the notion of mutual coherence between two tight frames $\Phi = (\phi_i)_i$ and $\Psi = (\psi_j)_j$,

$$\mu(\Phi, \Psi) = \max_j \max_i |\langle \phi_i, \psi_j \rangle|,$$

whose importance was shown by [6], as a means to impose conditions on the interactions between the dictionary elements. However, this notion is too weak for our purposes. Our novel contribution to sparse recovery and ℓ_1 minimization consists in exploiting the facts that

- the nonzeros of sparse vectors often do not arise in arbitrary patterns, but are rather highly structured, and that
- the interactions between the dictionary elements in ill-posed problems are not arbitrary, but rather geometrically driven.

These key observations lead to the following new notion.

Definition 1. *Given tight frames $\Phi = (\phi_i)_i$ and $\Psi = (\psi_j)_j$ and an index subset $\mathcal{S}$ associated with expansions in frame Φ, we define the* **cluster coherence**

$$\mu_c(\mathcal{S}; \Phi, \Psi) = \max_j \sum_{i \in \mathcal{S}} |\langle \phi_i, \psi_j \rangle|.$$

Thus cluster coherence bounds between a single member of frame Ψ and a cluster of members of frame Φ, clustered at $\mathcal{S}$, in contrast to mutual coherence, which can be thought of as singleton coherence.
A related notion called 'cumulative coherence' was introduced in [16], but notice that here we fix a specific set of significant coefficients and do not maximize over all such subsets. The key idea for our analysis is that the index subsets we consider are not abstract, but have a specific geometric interpretation. Maximizing over all subsets with a common combinatorial property would prohibit utilizing this interpretation, hence cumulative coherence is not suitable for our purposes.

2.2 Component Separation by ℓ_1 Minimization

Now consider the following optimization problem:

$$(\text{SEP}) \quad (S_1^\star, S_2^\star) = \operatorname{argmin}_{S_1, S_2} \|\Phi_1^T S_1\|_1 + \|\Phi_2^T S_2\|_1 \quad \text{subject to } S = S_1 + S_2.$$

Notice that in this problem, the norm is placed on the **analysis** coefficients rather than on the **synthesis** coefficients as in (BP) to avoid 'self-terms' in the frame expansions. The introduction of cluster coherence now ensures that the principle (SEP) gives a successful approximate separation.

Proposition 1 ([7]). *Suppose that S can be decomposed as $S = S_1^0 + S_2^0$ so that each component S_i^0 is relatively sparse in Φ_i, $i = 1, 2$, i.e.,*

$$\|1_{\mathcal{S}_1^c} \Phi_1^T S_1^0\|_1 + \|1_{\mathcal{S}_2^c} \Phi_2^T S_2^0\|_1 \leq \delta.$$

*Let $(S_1^\star, S_2^\star)$ solve (*SEP*). Then*

$$\|S_1^\star - S_1^0\|_2 + \|S_2^\star - S_2^0\|_2 \leq \frac{2\delta}{1 - 2\mu_c},$$

where

$$\mu_c = \max(\mu_c(\mathcal{S}_1; \Phi_1, \Phi_2), \mu_c(\mathcal{S}_2; \Phi_2, \Phi_1)).$$

3. Geometric Separation of Pointlike and Curvelike Structures

3.1 Radial Wavelet-Curvelet Dictionary

The concepts of the previous section will now be applied to $S = f_j^C = \mathcal{P}_j^C + \mathcal{C}_j^C$, our signal of interest from (2). The tight frames are Φ_1, the full radial wavelet frame, and Φ_2, the full curvelet tight frame. The subsignals $S_1^\star, S_2^\star$ we derive by applying the optimization problem (SEP) will be relabel to W_j, the wavelet component, and C_j, the curvelet component.
The main difficulty in applying Proposition 1 consists in choosing the sets of significant coefficients suitably. We achieve this by using microlocal analysis to understand heuristically the location of the significant coefficients in phase space. Roughly speaking, we then employ the Hart-Smith phase space metric defined by

$$\begin{aligned} d((b,\theta);(b',\theta')) &= |\langle e_\theta, b - b' \rangle| + |\langle e_{\theta'}, b - b' \rangle| \\ &\quad + |b - b'|^2 + |\theta - \theta'|^2 \end{aligned}$$

to define an 'approximate' set of significant wavelet coefficients

$$\begin{aligned} \Lambda_{1,j} &= \{\text{wavelet lattice}\} \\ &\quad \cap \{(b,\theta) : d((b,\theta); WF(\mathcal{P})) \leq \eta_j a_j\} \end{aligned}$$

and an 'approximate' set of significant curvelet coefficients

$$\begin{aligned} \Lambda_{2,j} &= \{\text{curvelet lattice}\} \\ &\quad \cap \{(b,\theta) : d((b,\theta); WF(\mathcal{C})) \leq \eta_j a_j\} \end{aligned}$$

for carefully chosen η_j; WF denotes the wavefront set.
Tedious, highly technical estimates then lead to the following separation result:

Theorem 1 ([7]). ASYMPTOTIC SEPARATION USING A RADIAL WAVELET-CURVELET DICTIONARY.

$$\frac{\|W_j - \mathcal{P}_j^C\|_2 + \|C_j - \mathcal{C}_j^C\|_2}{\|\mathcal{P}_j^C\|_2 + \|\mathcal{C}_j^C\|_2} \to 0, \qquad j \to \infty.$$

This result shows that components are recovered asymptotically: at fine scales, the energy in the curvelike component is all captured by the curvelet coefficients and the energy in the pointlike component is all captured by the wavelet coefficients.

3.2 Sparsity Equivalence

We now aim to show that curvelets and shearlets are *sparsity equivalent* in the sense that, for $0 < p \leq 1$, the ℓ_p norm of the curvelet coefficient sequence is finite if and only if the same is true for the shearlet coefficient sequence.
First we observe that for two tight frames $\Phi = (\phi_i)_i$ and $\Psi = (\psi_j)_j$, their cross-Grammian matrix

$$M(i,j) = \langle \phi_i, \psi_j \rangle$$

contains all information on the relation between coefficient sequences $\Phi^T S$ and $\Psi^T S$ for some signal S. Sparsity equivalence can therefore be proven by analyzing the p-norm, $0 < p \leq 1$ defined by

$$\begin{aligned} \|M\|_p &= \max\Big((\sup_i \sum_j |M(i,j)|^p)^{1/p}, \\ &\qquad (\sup_j \sum_i |M(i,j)|^p)^{1/p}\Big) \end{aligned}$$

of a cross-Grammian matrix M.
Now setting $(\sigma_\eta)_\eta$ to be the shearlet tight frame and $(\gamma_\mu)_\mu$ to be the curvelet tight frame, we derive the following result. We remark that the low frequency part has to be dealt with particular care, but for these technicalities we refer to [7].

Proposition 2 ([8]). *For all* $0 < p \leq 1$,

$$\|(\langle \sigma_\eta, \gamma_\mu \rangle)_{\eta,\mu}\|_p < \infty.$$

Using basic estimates from frame theory and the previous proposition, we can show that shearlets and curvelets are indeed sparsity equivalent, thereby allowing us to easily transfer results about sparsity from one system to the other.

Theorem 2 ([8]). *Let* $f \in L^2(\mathbf{R}^2)$ *and* $0 < p \leq 1$. *Then* $\|(\langle f, \sigma_\eta \rangle)_\eta\|_p < \infty$ *if and only if* $\|(\langle f, \gamma_\mu \rangle)_\mu\|_p < \infty$.

3.3 Orthonormal Wavelet-Shearlet Dictionary

Similar to Subsection 3.1, $S = f_j^S = \mathcal{P}_j^S + \mathcal{C}_j^S$ (see (2)) is now our signal of interest, and the tight frames are Φ_1, the full orthonormal wavelet frame, and Φ_2 the full shearlet tight frame. The subsignals $S_1^\star, S_2^\star$, we derive by applying the optimization problem (SEP) will be relabel to W_j, the wavelet component, and S_j, the shearlet component.
The results from Subsection 3.2 as well as similar correspondences between radial wavelets and orthonormal wavelets now form the backbone for the transfer of Theorem 1 to the orthonormal wavelet-shearlet dictionary. Careful application of those to the key estimates in the proof of Theorem 1 leads to a similar result for the orthonormal wavelet-shearlet dictionary.

Theorem 3 ([7]). ASYMPTOTIC SEPARATION USING AN ORTHONORMAL WAVELET-SHEARLET DICTIONARY.

$$\frac{\|W_j - \mathcal{P}_j^S\|_2 + \|S_j - \mathcal{C}_j^S\|_2}{\|\mathcal{P}_j^S\|_2 + \|\mathcal{C}_j^S\|_2} \to 0, \qquad j \to \infty.$$

4. Conclusion

We first considered signals, being a superposition of two subsignals, each of which is relatively sparse with respect to some tight frame. As a model procedure for separation we considered ℓ_1 minimization of the analysis (rather than synthesis) frame coefficients. By introducing cluster coherence as a new concept for analyzing the interaction of the two tight frames by taking the geometry of the sparse component expansions into account, we derived an estimate for the ℓ_2 norm of the separation error. We then considered signals, which are a superposition of pointlike and curvelike structures. Using the previously derived estimate, we proved that for both pairs of tight frames (radial wavelets/curvelets) as well as (orthonormal wavelets/shearlets) at sufficiently fine scale, nearly-perfect separation is achieved using the model procedure, thereby proposing the orthonormal wavelet-shearlet dictionary as an interesting alternative for geometric separation of pointlike and curvelike structures. The sparsity equivalence between curvelets and shearlets we further proved thereby allows to derive this separation result only for one dictionary and easily transfer it to the other one.

Acknowledgment

The authors would like to thank Emmanuel Candès, Michael Elad, and Jean-Luc Starck, for numerous discussions on related topics. The second author would like to thank the Department of Statistics at Stanford University and the Department of Mathematics at Yale University for their hospitality and support during her long-term visits. The authors would also like to thank the Newton Institute of Mathematics in Cambridge, UK for providing an inspiring research environment which led to the completion of a significant part of this work during their stay. This work was partially supported by NSF DMS 05-05303 and DMS 01-40698 (FRG), and by Deutsche Forschungsgemeinschaft (DFG) Heisenberg Fellowship KU 1446/8-1. We further thank the anonymous referee for useful comments and suggestions.

References:

[1] A.M. Bruckstein, D.L. Donoho, and M. Elad. From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images. *SIAM Review* 51:34–81, 2009.

[2] E. J. Candès and D. L. Donoho. Continuous curvelet transform: I. Resolution of the wavefront set. *Appl. Comput. Harmon. Anal.* 19:162–197, 2005.

[3] E. J. Candès and D. L. Donoho. Continuous curvelet transform: II. Discretization of frames. *Appl. Comput. Harmon. Anal.* 19:198–222, 2005.

[4] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review* 43:129–159, 2001.

[5] R. R. Coifman and M. V. Wickerhauser. Wavelets and adapted waveform analysis. A toolkit for signal processing and numerical analysis, In *Different*

perspectives on wavelets (San Antonio, TX, 1993), 47:119–153, Proc. Sympos. Appl. Math., Amer. Math. Soc., Providence, RI, 1993.

[6] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory* 47:2845–2862, 2001.

[7] D. L. Donoho and G. Kutyniok. Microlocal Analysis of the Geometric Separation Problem. Preprint, 2009.

[8] D. L. Donoho and G. Kutyniok. Sparsity Equivalence of Anisotropic Decompositions. Preprint, 2009.

[9] M. Elad, J.-L. Starck, P. Querre, and D. L. Donoho. Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA). *Appl. Comput. Harmon. Anal.* 19:340–358, 2005.

[10] R. Gribonval and M. Nielsen. Sparse representations in unions of bases. *IEEE Trans. Inform. Theory* 49:3320–3325, 2003.

[11] K. Guo, G. Kutyniok, and D. Labate. Sparse Multidimensional Representations using Anisotropic Dilation und Shear Operators. In *Wavelets und Splines* (Athens, GA, 2005), G. Chen und M. J. Lai, eds., Nashboro Press, Nashville, TN (2006), 189–201.

[12] M. Kowalski and B. Torrésani. Sparsity and Persistence: mixed norms provide simple signal models with dependent coefficients. *Signal, Image and Video Processing*, to appear.

[13] G. Kutyniok and D. Labate. Resolution of the Wavefront Set using Continuous Shearlets. *Trans. Amer. Math. Soc.* 361:2719–2754, 2009.

[14] F. G. Meyer, A. Averbuch, and R. R. Coifman. Multilayered Image Representation: Application to Image Compression. *IEEE Trans. Image Proc.* 11:1072–1080, 2002.

[15] J.-L. Starck, M. Elad, and D. L. Donoho. Image decomposition via the combination of sparse representations and a variational approach. *IEEE Trans. Image Proc.* 14:1570–1582, 2005.

[16] J. A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory* 50:2231–2242, 2004.

Special session on

Sampling and Communication

Chair: Götz PFANDER

A Kashin Approach to the Capacity of the Discrete Amplitude Constrained Gaussian Channel

Brendan Farrell [(1)] and Peter Jung [(2)]

(1) Heinrich-Hertz Lehrstuhl, Technische Universität Berlin, Einsteinufer 25, 10587 Berlin, Germany.
(2) Fraunhofer German-Sino Lab for Mobile Communications - MCI, Einsteinufer 37, 10587 Berlin, Germany.
brendan.farrell@mk.tu-berlin.de, peter.jung@hhi.fhg.de

Abstract:

We derive an explicit lower bound on the capacity of the discrete amplitude–constrained Gaussian channel by proving the existence of tight frames that permit redundant vector representations with small coefficients. Our method encodes the information in subspaces that are optimal in terms of the power to amplitude ratio. In a recent paper, Lyubarskii and Vershynin discuss how the work of Kashin (1977) implies the existence of such representations, and they term them Kashin respresentations. We use this work from frame theory to address the relationship between signal redundancy, peak–to–average power ratio and achievable data rates.

1. Introduction

Communication at high data rates and with moderate cost on hardware and complexity provide challenging topics in engineering and applied mathematics. An important problem in this direction is efficient signaling and coding under an amplitude constraint. In general, the cost for high data rate is related to a power budget. However, in practical communication systems, there sometimes exist disruptive or non-linear effects that only occur at high signal amplitudes. The information–theoretic treatment of amplitude–constrained channel is completely different from the power–constrained channel. On the other hand, coding for power–constrained Gaussian channels is well understood. Clearly, if a loss in data rate is accepted, signals can be constructed with lower maximum amplitude. The optimal scaling between power and amplitude and an explicit relation to achievable rates will be given in this paper. In this case, the data-rate loss is caused by considering redundant representations. Here, the original vectors are expanded with respect to a particular frame and the coefficients are then transmitted.

We show that there exist frames which allow the standard coding approach to be used for the amplitude-constrained channel. Our result is Theorem 2, which comes at the end of the paper. This theorem states that for the amplitude constrained, Gaussian channel the rate

$$\frac{1}{2\lambda_{\min}} \log\left(1+\lambda_{\min}\frac{\text{Signal Power}}{\text{Noise Power}}\right) \tag{1}$$

is achievable for a redundancy $\lambda_{\min}$ that is an *explicit function* of the peak-to-average power ratio. We note that by making the amplitude constraint compatible with Gaussian codebooks, we make the developed tools and understanding of Gaussian codebooks applicable to the amplitude–constrained channel. Results from frame theory, thus, allow us to address a question in information theory. While the results used from functional analysis are well known there, we show a new application.

1.1 The Information–Theoretic Problem

The capacity of a communication channel is the maximum amount of information per unit of time that can be sent from a sender through the channel to the receiver. Shannon made this operational concept mathematically rigorous by formulating it in terms of entropy [7]. In [7] Shannon addressed the discrete–time model:

$$Y = X + Z, \tag{2}$$

for the noisy channel, where X and Y denote the (real) channel input and output, and the additive noise Z is a Gaussian random variable with variance σ^2. Let X^n be a random vector in $\mathbb{R}^n$ according to a distribution to be determined and Z^n the random vector having n identical independent distributed (iid) copies of Z. Shannon introduced two concepts of a capacity for this model. The *information capacity* $C^{(i)}$ is the supremum of the information rates:

$$C^{(i)} = \lim_{n\to\infty} \frac{1}{n} \sup_{\mu^n \in \mathcal{F}^n} \mathcal{I}(X^n;Y^n) \tag{3}$$

taken over all distributions μ^n of X^n from a particular subset $\mathcal{F}^n \subset \mathcal{P}^n$ of probability distributions $\mathcal{P}^n$. $\mathcal{I}(X^n;Y^n)$ denotes the mutual information between the random variables X^n and Y^n and is equal to the entropy of Y^n minus the entropy of Y^n given X^n, $\mathcal{I}(X^n;Y^n) = h(Y^n) - h(Y^n|X^n)$. From its concavity in μ^n it follows that the optimum μ^n_{opt} is at least achieved for a product distribution, i.e. single letter coding with a measure $\mu = \mu^1$ is optimal in this sense. Shannon considered an averaged power constraint P which corresponds to the set $\mathcal{F} = \mathcal{F}^1$ of single–letter distributions:

$$\mathcal{F} = \{\mu \in \mathcal{P} \mid \int |x|^2 d\mu(x) \le P\} \tag{4}$$

or equivalently

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}|x_i|^2 \le P. \tag{5}$$

He found that the optimum μ_{opt} is attained for a Gaussian distribution with variance P and that

$$C^{(i)} = \frac{1}{2}\log(1+\frac{P}{\sigma^2}). \tag{6}$$

Shannon further showed with a so called *coding theorem* that it is even possible to get arbitrary close to that value justifying the term *channel capacity*. That is, for each rate $R < C^{(i)}$ there exist 2^{nR} codewords $\{X(\omega)\}_{\omega=1}^{\omega=2^{nR}}$ in $\mathbb{R}^n$ (called a $(2^{nR}, n)$ code) such that $X(\omega) + Z^n$ can be distinguished at the receiver with error probability going to zero as n increases. (X will now denote codewords and be indexed by ω.) Each admissible codeword satisfies the average power constraint $\frac{1}{n}\sum_{i=1}^{n}|X_i(\omega)|^2 \leq P$; however, to achieve the capacity it may be necessary to use codewords having maximum amplitudes which scale with $\sqrt{n}$.

We address an additive, white Gaussian noise (AWGN) channel under the assumption that there is both a power constraint,

$$\frac{1}{n}\sum_{i=1}^{n}|X_i(\omega)|^2 \leq P, \tag{7}$$

and a strict amplitude constraint:

$$\max_{i=1,\ldots,n}|X_i(\omega)| \leq A, \tag{8}$$

for two positive, real numbers P and A and for all $\omega = 1, ..., 2^{nR}$.

The information capacity under a constraint A on the amplitudes of the signals was solved by Smith [8]. Similar to the Gaussian channel with power constraint only, Smith showed that the capacity of the amplitude–constrained channel is attained when the entries x_i are independent.

The set of (single–letter) input distributions is in this case:

$$\mathcal{F} = \{\mu \in \mathcal{P} \,|\, \mu(\{|x| > A\}) = 0\}. \tag{9}$$

Smith found that the optimum measure μ_{opt} has discrete and finite support. Similar results are known for other noise densities (see for example [6]). A characterization of the number of mass points in the Gaussian case is unknown. For a given assumption on this number the values and the positions can be computed. From this Smith gave an algorithm which numerically computes $C^{(i)}$. Smith establishes an algorithm to determine the optimal input probability measure given the constraints A, P and σ^2. However, to date there is not a general strategy applicable for a practical range of these parameters.

1.2 Frames and Banach Geometry

We will work strictly with real numbers. We have the following norms for $\mathbb{R}^n$: $\|x\|_{l_p^n} = (\sum_{i=1}^{n}|x_i|^p)^{1/p}$ and $\|x\|_{l_\infty^n} = \max_{i=1,\ldots,n}|x_i|$. B_p^n will denote the unit ball in $\mathbb{R}^n$ with respect to the ℓ_p-norm. We denote by U_n^N an n-dimensional subspace of $\mathbb{R}^N$, $N \geq n$. We will often speak of a matrix $U \in \mathbb{R}^{n\times N}$ whose rows are orthonormal and span U_n^N or whose columns constitute a tight frame for $\mathbb{R}^n$.

Definition 1. A set of vectors $\{u_i\}_{i=1}^N \subset \mathbb{R}^n$ is a *tight frame* for $\mathbb{R}^n$ if

$$\|x\|_2^2 = \sum_{i=1}^{N} = |\langle x, u_i\rangle|^2 \tag{10}$$

for all $x \in \mathbb{R}^n$.

It follows that the columns of an $n\times N$ matrix U constitute a tight frame for $\mathbb{R}^n$ if and only if $UU^* = I_n$, where I_n denotes the identity matrix of size n. In the proof of the coding theorem (see, for example, [1]) for the Gaussian channel with average power constraint P, the constructed codewords $X \in \mathbb{R}^n$ satisfy the constraint $\|X\|_{\ell_2^n} \leq \sqrt{nP}$. Similarly, in the amplitude constrained channel codewords must satisfy $\|X\|_{\ell_\infty^n} \leq A$. In other words, admissible signals X for the amplitude constraint channel lie in a scaled cube, i.e. $X \in A \cdot B_\infty^n$. And for a power constrained channel the signals are contained in an increasing ball $X \in \sqrt{nP} \cdot B_2^n$.

Of course the difficult aspect of this channel is the amplitude constraint. We do not require that the random input variables $\{x_i\}_{i=1}^n$ be independent, which allows us to use redundant representations.

The basic idea for our approach is the following: given N vectors $\{u_i\}_{i=1}^N$ spanning $\mathbb{R}^n$, $N > n$, a vector $x \in \mathbb{R}^n$ may be expressed, in general, in multiple ways as a linear combination of the vectors $\{u_i\}_{i=1}^N$:

$$x = \sum_{i=1}^{N} b_i u_i. \tag{11}$$

In light of the amplitude constraint, the question is whether one of the possible expressions (10) satisfies $\|b\|_{l_\infty^N} \leq A$. If this is possible, then we may transmit the vector b and suffer an efficiency loss of $N - n$ symbols.

The representation (10) is called a *Kashin representation* [5] of the vector x if $\|b\|_\infty \leq C\|x\|_2$. We first address a general frame setting and then focus on the Kashin representations in Section 3.

2. General Frame Setting

As we have seen, the capacity of the discrete Gaussian channel with average power constraint P and noise variance σ^2 is $\frac{1}{2}\log(1+\frac{P}{\sigma^2})$. This means, If $R < \frac{1}{2}\log(1+\frac{P}{\sigma^2})$ is the rate, then there are 2^{nR} codewords, and all admissible codewords for this channel satisfy the power constraint $\|X(\omega)\|_2 \leq \sqrt{nP}$, $\omega = 1, ..., 2^{nR}$. If one has a tight frame $\{u_i\}_{i=1}^N$ for $\mathbb{R}^n$, $N = [\lambda n]$, then one can also achieve the rate:

$$\frac{1}{2\lambda}\log(1+\frac{\lambda P}{\sigma^2}) \tag{12}$$

by transmitting codewords $\{Y(\omega)\}_{\omega=1}^{2^{nR}} \subset \mathbb{R}^N$ satisfying $UY(\omega) = X(\omega)$ for $\omega = 1, ..., 2^{nR}$.

Since columns of $U \in \mathbb{C}^{n\times N}$ form a tight frame for $\mathbb{R}^n$, $\|UX(\omega)\|_{l_N^2} = \|Y(\omega)\|_{l_n^2}$, and thus:

$$\frac{1}{N}\sum_{i=1}^{N}|Y_i(\omega)|^2 = \frac{1}{\lambda n}\sum_{i=1}^{n}|X_i(\omega)|^2 \leq P. \tag{13}$$

The key point is that a vector $Y(\omega)$ that satisfies $UY(\omega) = X(\omega)$ is, in general, not unique. For a given additional constraint, one may ask if there exists a set $\mathcal{Y} \subset \mathbb{R}^N$ satisfying the additional constraint and a tight frame with matrix U such that:

$$U\mathcal{Y} = \{x | x \in \mathbb{R}^n,\ \|x\|_2 = 1\}. \tag{14}$$

The existence of such a set and a corresponding tight frame is sufficient to imply that $\frac{1}{2\lambda}\log(1 + \frac{\lambda P}{\sigma^2})$ is an achievable rate for the discrete Gaussian channel with the additional constraint.

The additional constraint of interest here is the amplitude constraint; that is, it is required that $\|Y(\omega)\|_{l_\infty^N} \leq A$ for all codewords $Y(\omega)$. Thus, for a given codebook $\{X(\omega)\}_{\omega=1}^{2^{nR}}$ satisfying $\|X(\omega)\|_{l_2^n} \leq \sqrt{nP}$ for all ω, we would like to determine a second codebook $\{Y(\omega)\}_{\omega=1}^{2^{nR}} \subset \mathbb{R}^N$ satisfying $\|Y(\omega)\|_{l_\infty^N} \leq A$ and a tight frame so that $UY(\omega) = X(\omega)$ for all ω. For completeness and clarity, we include the communication strategy. The next section will show that Step 2 is possible for an appropriate λ.

Communication Strategy:

1. The set of vectors $\{u_i\}_{i=1}^N$ form a tight frame for $\mathbb{R}^n$ and are known to both transmitter and receiver.
2. Each codeword $X(\omega)$ satisfies the power constraint , and its Kashin representation $Y(\omega) \in \mathbb{R}^N$ satisfying $\|Y(\omega)\|_{l_\infty^N} \leq A$ is determined.
3. To transmit the message ω, the transmitter sends $Y(\omega)$.
4. $Y(\omega) + Z^N \in \mathbb{R}^N$ is received.
5. Receiver multiplies $Y(\omega) + Z^N$ by U to obtain $X(\omega) + UZ^N \in \mathbb{R}^n$.
6. Receiver decodes $X(\omega) + UZ^N \in \mathbb{R}^n$.

We note that, in contrast to the approach of Smith [8], this approach is still based on Gaussian codebooks, and, therefore, the extensive tools developed for Gaussian codebooks are still applicable.

3. Kashin Representations or Optimal Subspaces

Definition 2 (Kashin Representations)**.** For a set of vectors $\{u_i\}_{i=1}^N \subset \mathbb{R}^n$, $N > n$, the expansion

$$x = \sum_{i=1}^{N} a_i u_i \tag{15}$$

is a *Kashin representation with level* K of the vector $x \in \mathbb{R}^n$ if

$$\|a\|_{l_\infty^N} \leq \frac{K\|x\|_{l_2^n}}{\sqrt{N}}, \quad i = 1, ..., N. \tag{16}$$

See [3, 4, 5]. We denote by U the $n \times N$ dimensional matrix with columns $\{u_i\}_{i=1}^N$. If these vectors constitute a tight frame, then $UU^* = I_n$, where I_n denotes the identity matrix on $\mathbb{R}^n$. One possible coefficient vector for equation (14) is $a = U^*x$. For this vector, we note

$$\|a\|_{l_\infty^N} \leq \|a\|_{l_2^N} = \sqrt{\langle U^*x, U^*x\rangle} \tag{17}$$

$$= \sqrt{\langle I_n x, x\rangle} = \|x\|_{l_2^n}. \tag{18}$$

Consequently, for a tight frame, it is always possible to find a vector a satisfying $\|a\|_{l_\infty^N} \leq \|x\|_{l_2^n}$, and thus equation (15) can be satisfied for every tight frame with Kashin level $K = \sqrt{N}$.

Of course the study of Kashin representations is concerned with optimally small constants and their relation to the redundancy $\lambda = N/n$. We will be interested in the dependence of $K = K(\lambda)$ on λ, but we postpone the discussion of the constant $K(\lambda)$ until the next section. Now, we show a lower bound on the achievable capacity when the amplitude constraint is $K(\lambda)\sqrt{P}$ (or greater).

If we set any n orthonormal vectors in $\mathbb{R}^N$ to be the rows of a matrix U, then $UU^* = I_n$, and the columns of U constitute a tight frame for $\mathbb{R}^n$. Thus, a tight frame for $\mathbb{R}^n$ can be constructed from any n-dimensional subspace of $\mathbb{R}^N$. For $U \in \mathbb{C}^{n\times N}$, let U_n^N denote the subspace of $\mathbb{R}^N$ spanned by its rows. Then $U(B_2^N \cap U_n^N) = B_2^n$. Therefore, for any $x \in B_2^n$, as long as the rows of U are linearly independent there exists a $y \in (B_2^N \cap U_n^N)$ such that $x = Uy$. In the higher dimensional space, we have an $\|\cdot\|_{l_\infty^N}$-norm constraint. We thus want to find an n-dimensional subspace of $\mathbb{R}^N$ that can be mapped isometrically with respect to the $\|\cdot\|_{l_2^n}$-norm to $\mathbb{R}^n$, and we must be able to cover B_2^n in this way.

First results on the smallest constant C, such that a projection of the ball $C \cdot B_\infty^N$ covers B_2^n was given by Kashin in [3]. There he showed that the scaling is $\mathcal{O}(n^{-1/2})$, and the exact optimal scaling was then determined in [2]. Since the $\|\cdot\|_2$-isometric projection is equivalent to the existence of a tight frame, we formulate their result in terms of frames.

Theorem 1 ([3, 2])**.** *For all positive integers N and n, $N > n$, there exists a tight frame for $\mathbb{R}^n$ consisting of N vectors such that every vector in $\mathbb{R}^n$ has a Kashin representation of level:*

$$K(\lambda) := C\left(\frac{\lambda}{\lambda - 1}\log\left(1 + \frac{\lambda}{\lambda - 1}\right)\right)^{1/2}, \tag{19}$$

where $\lambda = N/n$ with respect to this frame.

See also [4, 5] for further discussion of this result. In [5] Lyubarskii and Vershynin have recently given an algorithm for determining a Kashin representation. In the same paper they discuss various ways to generate the required frames and determine their Kashin constants.

Theorem 2. *For a given amplitude constraint A, there exists a constant $\lambda_{\min}$ such that the capacity $\mathcal{C}_{P,A}$ of the discrete Gaussian channel with average power constraint P, amplitude constraint A and noise variance σ^2 is lower bounded by*

$$\mathcal{C}_{P,A} \geq \frac{1}{2\lambda_{\min}}\log\left(1 + \frac{\lambda_{\min}P}{\sigma^2}\right). \tag{20}$$

Proof Theorem 1 shows the existence of a frame with the necessary properties, as discussed in the communication strategy in Section 2. Denoting the matrix corresponding to this frame by U, for each codeword $X(\omega) \in \mathbb{R}^n$, there exists a codeword $Y(\omega) \in \mathbb{R}^N$ such that $X(\omega) = UY(\omega)$, and

$$\begin{aligned} \|Y(\omega)\|_{l_\infty^N} &\leq \frac{K(\lambda_{\min})}{\sqrt{N}} \|X(\omega)\|_{l_2^n} \qquad (21) \\ &\leq K(\lambda_{\min})\sqrt{P}. \qquad (22) \end{aligned}$$

Lastly, $\lambda_{\min}$ is the solution to

$$C\left(\frac{\lambda}{\lambda-1}\log(1+\frac{\lambda}{\lambda-1})\right)^{1/2} = \frac{A}{\sqrt{P}}, \qquad (23)$$

which exists and is unique since $\left(\frac{\lambda}{\lambda-1}\log(1+\frac{\lambda}{\lambda-1})\right)^{1/2}$ is monotone increasing. ❏

4. Conclusion

We have considered an application of the redundant representations found in frame theory and geometric functional analysis to a fundamental question in information theory.

References:

[1] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[2] A. Garnaev and E. D. Gluskin. The widths of euclidean balls. *Doklady An. SSSR.*, 277:1048–1052, 1984.

[3] B. S. Kashin. Diameters of some finite-dimensional sets and classes of smooth functions. *Izv. Akad. Nauk SSSR Ser. Mat.*, 41(2):334–351, 478, 1977. English transl. in Math. USSR IZV. 11 (1978), 317-333.

[4] B. S. Kashin and V. N. Temlyakov. A remark on compressed sensing. *Mathematical Notes*, 82(5):748–755, Nov 2007.

[5] Y Lyubarskii and R. Vershynin. Uncertainty principles and vector quantization. preprint.

[6] W. Oettli. Capacity-achieving input distributions for some amplitude-limited channels with additive noise (corresp.). *IEEE Transactions on Information Theory*, 20(3):372–374, May 1974.

[7] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423,623–656, 1948.

[8] Joel G. Smith. The Information Capacity of Amplitude and Variance Constrained Scalar Gaussian Channels. *Information and Control*, 18:203–219, 1971.

Erasure-proof coding with fusion frames

Bernhard G. Bodmann[(1)], Gitta Kutyniok[(2)] and Ali Pezeshki[(3)]
(1) Department of Mathematics, University of Houston, Houston, TX 77204, USA
(2) Institute of Mathematics, University Osnabrueck, 49069 Osnabrueck, Germany
(3) Electrical and Computer Engineering Department, Colorado State University, Fort Collins, CO 80523, USA
bgb@math.uh.edu, kutyniok@uni-osnabrueck.de, pezeshki@engr.colostate.edu

Abstract:

The main goal of this paper is the design of frames for transmitting vectors through a memoryless analog erasure channel. The channel transmits the frame coefficients perfectly or discards them, depending on the outcomes of Bernoulli trials with a failure probability q. For sufficiently small q, we construct frames which encode above a fixed non-zero rate and allow the receiver to recover part of the erased coefficients so that the remaining mean-square error vanishes as the frame size increases. We give examples for which the mean-square reconstruction error remaining after corrections are applied decays faster than any inverse power of the number of frame vectors.

1. Introduction

We are concerned with the linear transmission of vectors through a memoryless channel that either transmits a coefficient perfectly or discards it, in accordance with the outcomes of independent, identically distributed Bernoulli trials. The problem of reconstructing a vector in a finite-dimensional real or complex Hilbert space when not all of its frame coefficients are known has already received much attention in the literature [1–9]. However, many results focus on optimal performance for the smallest possible number of erased coefficients [4, 7–9], which is not typical for transmissions via a memoryless erasure channel. Other results on so-called maximally robust frames guarantee recovery from a certain fraction of lost frame coefficients [10], but this may involve inverting an arbitrarily ill-conditioned matrix.

The notion of a memoryless analog erasure channel is simply one that transmits each frame coefficient independently with a given success probability q and otherwise erases it, meaning it does not let the receiver access the coefficient. Within this error model for transmissions, we investigate the performance of fusion frames [11–13], previously also referred to as frames of subspaces [14] or weighted projections resolving the identity [15], which lend themselves to various methods of error correction. What makes the fusion frames useful for error correction purposes is that they have many subsets which are frames for their span. Thus, one can design hierarchical methods for error correction which make error estimates feasible.

The main result presented here is that for a fixed, sufficiently small erasure probability q, we design fusion frames such that their associated coding rate is bounded away from zero and the mean-square error remaining after error correction is applied decays faster than any polynomial in terms of the number of frame vectors.

The techniques for our results involve combinatorial elements similar to the construction of product codes initially investigated by Elias [16], together with some frame-specific arguments.

2. Preliminaries

Throughout the paper, we let $\mathcal{H}$ be a real or complex Hilbert space. Instead of expanding vectors in Hilbert spaces with orthonormal bases, many applications nowadays use frames, stable, non-unique (redundant) expansions, for various purposes. We first briefly recall the basic terminology, and refer the reader to [17] for further details.

Definition 1. *We call a family of vectors $\mathcal{F} = \{f_j\}_{j\in J}$ in $\mathcal{H}$ a* frame *if there exist constants $A, B > 0$ such that for all $x \in \mathcal{H}$ with $\|x\| = 1$, $A \leq \sum_{j\in J} |\langle x, f_j\rangle|^2 \leq B$. If we can choose $A = B$, then we say that the frame is* A-tight. *In case $A = B = 1$ we call $\mathcal{F}$ a* Parseval frame. *A frame is called* equal-norm *if there is a $c > 0$ such that all vectors have the norm $\|f_j\| = c$. With each frame $\mathcal{F}$, we associate the* analysis operator $V : \mathcal{H} \to \ell^2(J)$, *which maps a vector to its frame coefficients, $(Vx)_j = \langle x, f_j\rangle$.*

The fact that a vector is over-determined by its frame coefficients helps correct errors which may occur in the course of a transmission, or when frame coefficients are stored in an unreliable medium. A main goal of frame design is to optimize the performance of a frame given certain constraints. This could be, for example, the dimension of the Hilbert space and the number of frame vectors, or their ratio. In analogy with binary codes, we define a coding rate for a given frame.

Definition 2. *Let $\mathcal{H}$ be a Hilbert space of dimension d and $\mathcal{F}$ a frame for $\mathcal{H}$ consisting of n vectors. We say that $\mathcal{F}$ has a* coding rate *of $R = d/n$.*

The coding and error correction method we discuss hereafter relies on frames arising from tensor product constructions. These frames are a special type of a fusion frame, see e.g. [12–15].

Definition 3. *Given Hilbert spaces $\mathcal{H}_1, \mathcal{H}_2, \dots \mathcal{H}_m$ and tight frames $\mathcal{F}^{(i)} = \{f_j^{(i)}\}_{j\in J_i}$ for each $\mathcal{H}_i$, then the family of vectors $\mathcal{F} = \{f_{j_1}^{(1)} \otimes f_{j_2}^{(2)} \otimes \cdots \otimes f_{j_m}^{(m)} : j_i \in J_i \text{ for all } i\}$ is a tight frame for $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \cdots \otimes \mathcal{H}_m$. We call this frame $\mathcal{F}$ a tight* product frame.

Remark 1. *We note that if we fix all but one index, say the last, then the resulting set $f_{j_1}^{(1)} \otimes f_{j_2}^{(2)} \otimes \cdots \otimes f_{j_{m-1}}^{(m-1)} \otimes \mathcal{F}^{(m)}$ is a tight frame for its span. Therefore, $\mathcal{F}$ has a natural fusion frame architecture.*
Similarly, fixing only the first $m-k$ indices of the frame vectors in the tensor product would provide a tight frame for a subspace for any $0 \le k < m$. Moreover, there is a partial ordering on these tight frames for subspaces induced by the partial ordering of the subspaces they span.

3. Erasures and the mean-square error

A communication system is given by a frame $\mathcal{F}$ for a Hilbert space $\mathcal{H}$, and an error model for the transmission of frame coefficients. Our main error model assumes memoryless erasures, that is, the values of randomly selected frame coefficients become unknown in the course of transmission, in accordance with the outcomes of Bernoulli trials. In brief, frame coefficients are erased, independently of each other, with a fixed probability $q \ge 0$.
Depending on the implementation of decoding, the performance of a frame can be measured in different ways; we generally distinguish active error correction and blind reconstruction. When actively correcting erasures, one tries to fill in the values for the erased coefficients, and aims for a high probability of successfully restoring all lost coefficients. When blind reconstruction is used, one sets the missing coefficients to zero and reconstructs always in the same way. In this case, the usual goal is obtaining a small error norm, such as the mean-square error or the worst-case error.
In the present work we consider a combination of the two approaches. We measure the quality of error correction by the mean-square error that results from using the corrected coefficients with the possibly remaining, uncorrected erasures set to zero. The average in this mean-square error is taken over the random erasures and over random unit-norm input vectors. For simplicity, we consider input vectors which are independent of the erasures and uniformly distributed on the unit sphere of the Hilbert space.

Definition 4. *Let $\mathcal{F} = \{f_1, f_2, \dots f_n\}$ be a Parseval frame for a real or complex Hilbert space $\mathcal{H}$. The* blind reconstruction error *for an input vector $x \in \mathcal{H}$ and an erasure of frame coefficients with indices $K = \{j_1, j_2, \dots j_m\}$, $m \le n$, is given by*

$$\|V^*EVx - x\| = \|(V^*EV - I)x\|$$

where E is the diagonal $n \times n$ matrix with $E_{j,j} = 1$ if $j \notin K$ and $E_{j,j} = 0$ else. If the positive operator V^*EV has a bounded inverse, then we say that the corresponding erasure is correctible.

Remark 1. *If $\mathcal{F}$ is a Parseval frame then $(V^*EV - I)x = V^*(E-I)Vx$ and the inverse can be obtained from the norm-convergent Neumann series $(V^*EV)^{-1} = \sum_{n=0}^{\infty}(V^*(I-E)V)^n$. Applying this operator to the output of blind reconstruction gives perfect reconstruction of the input vector.*

Next, we define a measure for average reconstruction performance when probabilities for erasures are known. To this end, we average the square of the reconstruction error with the distribution of erasures and input vectors. Here and herafter, we denote the expectation of any random variable η with respect to the underlying probability measure $\mathbb{P}$ by $\mathbb{E}[\eta] = \int \eta d\mathbb{P}$.

Definition 5. *Let $\{\beta_j\}_{j\in J}$ be a family of binary ($\{0,1\}$-valued) random variables governed by a probability measure $\mathbb{P}$, and let Δ be the random diagonal matrix with entries $\Delta_{j,j} = \beta_j$. Moreover, let ξ be a random variable with values in the unit sphere $\{x \in \mathcal{H} : \|x\| = 1\}$ which is independent of the family $\{\beta_j\}$, and assume that the distribution of $U\xi$ is identical to that of ξ for any fixed unitary U. Given a Parseval frame $\mathcal{F}$ for a Hilbert space $\mathcal{H}$ with analysis operator V, we define the* mean-square error *by*

$$\sigma^2(V,\beta) = \mathbb{E}[\|V^*\Delta V\xi\|^2]\,.$$

There is a simple expression for the mean square error as the square of a weighted Frobenius norm of the Grammian VV^*.

Lemma 1. *Let $\{\beta_j\}_{j\in J}$ be as above, assume the family is identically distributed with probability $\mathbb{P}(\beta_1 = 1) = q$, and assume the joint distribution is such that $\mathbb{P}(\beta_j = \beta_{j'} = 1) = r$ for all $j \ne j'$. Let Δ be the random diagonal matrix with entries $\Delta_{j,j} = \beta_j$. If V is the analysis operator of a Parseval frame $\mathcal{F} = \{f_j\}_{j\in J}$ containing $n = |J|$ vectors in a Hilbert space of dimension d, then*

$$\sigma^2(V,\beta) = \frac{1}{d}\Big((q-r)\sum_{j=1}^{n}\|f_j\|^4 + r\sum_{j,l=1}^{n}|\langle f_j, f_l\rangle|^2\Big)\,.$$

4. Bounding the mean-square error for iterative decoding

This section describes how product frames can be used to trade an increase in block length of encoding for better error correction capabilities.
We first consider the simplest case in which $\mathcal{H}$ has two factors, $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$. Also, as preparation for our main theorem, we first consider packet erasures [15] instead of erasures for single frame coefficients. This means, we have a frame $\mathcal{F} = \mathcal{F}^{(1)} \otimes \mathcal{F}^{(2)}$ and a two-parameter family of random variables $\{\beta_{j,j'}\}$ which govern erasures of frame coefficients in such a way that either all coefficients belonging to some j' are erased or all of them are left intact. We compute the mean-square error for this error model.

Proposition 1. *Let $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$ and let V_1 and V_2 be the analysis operators of Parseval frames $\mathcal{F}^{(1)} = \{f_j^{(1)}\}_{j\in J_1}$ and $\mathcal{F}^{(2)} = \{f_{j'}^{(2)}\}_{j'\in J_2}$ for $\mathcal{H}_1$ and $\mathcal{H}_2$ having dimension*

d_1 and d_2, respectively. Let $\{\beta_{j,j'} : j \in J_1, j' \in J_2\}$ be a two-parameter family of binary random variables which have probabilities $\mathbb{P}(\beta_{j,j'} = 1) = q$ and are distributed such that there is a family $\{\beta^{(2)}_{j'}\}_{j' \in J_2}$ and $\beta_{j,j'} = \beta^{(2)}_{j'}$ almost surely, regardless of j. The mean-square error for the frame $\mathcal{F}$ and this type of packet erasures reduces to that of $\mathcal{F}^{(2)}$,

$$\sigma^2(V_1 \otimes V_2, \beta) = \sigma^2(V_2, \beta^{(2)}) \,.$$

Next, we continue with three combinatorial lemmata. They prepare the main result which concerns the error correction capabilities of tight product frames. The main problem we wish to address with this result is the following: Given a fixed, sufficiently small erasure probability q, find frames such that their associated coding rate is bounded away from zero and the mean-square error remaining after error correction is applied decays fast in terms of the number of frame vectors.
We show hereafter that product frames of the form $\mathcal{F} = \mathcal{F}^{(1)} \otimes \cdots \otimes \mathcal{F}^{(m)}$, for which each factor $\mathcal{F}^{(i)}$ can correct up to two erased frame coefficients, satisfy the desired properties.

Lemma 2. *Let $n_1 \geq 3$ and let $\{\beta_1, \beta_2, \dots \beta_{n_1}\}$ be a family of independent, identically distributed random variables which take values in $\{0, 1\}$. Suppose $q_0 = \mathbb{P}(\beta_1 = 1)$ and let $q_1 = \mathbb{P}(\sum_{j=1}^{n_1} \beta_j \geq 3)$, then*

$$q_1 \leq \frac{1}{6} n_1^3 q_0^3 \,.$$

The probability estimated in this lemma is that of a packet of n_1 coefficients remaining corrupted after an error correction protocol has been applied which can correct any two erased coefficients.
By iteration, we obtain a simple consequence.

Lemma 3. *Let $\{n_i\}_{i=1}^m$ be the sizes of index sets $\{J_i\}_{i=1}^m$, with $n_i \geq 3$ for all $i \in \{1, 2, \dots m\}$. Assume there is an m-parameter family of binary, independent identically distributed random variables $\{\beta_{j_1,j_2,\dots j_m}\}$ and associated families $\{\beta^{(1)}_{j_2,j_3,\dots j_m}\}$, $\{\beta^{(2)}_{j_3,j_4,\dots j_m}\}$, $\dots \{\beta^{(m-1)}_{j_m}\}$ which are iteratively defined by $\beta^{(0)}_{j_1,j_2,\dots j_m} \equiv \beta_{j_1,j_2,\dots j_m}$ and*

$$\beta^{(k)}_{j_{k+1},j_{k+2},\dots j_m} = \begin{cases} 1, & \text{if } \sum_{j_k=1}^{n_k} \beta^{(k-1)}_{j_k,j_{k+1},\dots j_m} \geq 3 \,, \\ 0, & \text{else.} \end{cases}$$

If $\mathbb{P}(\beta_{1,1,\dots 1} = 1) = q_0$, then the family $\{\beta^{(m-1)}_j\}$ is independent, identically distributed with $q_{m-1} = \mathbb{P}(\beta^{(m-1)}_j = 1)$ having the bound

$$q_{m-1} \leq 6^{-\frac{1}{2}(3^{m-1}-1)} n_{m-1}^{3^1} n_{m-2}^{3^2} \cdots n_1^{3^{m-1}} q_0^{3^{m-1}} \,.$$

The probability computed in the above lemma is the probability of an erased block after applying erasure correction iteratively. The next lemma considers what happens when the error correction is applied to packets at the final level. Here, we deviate from the strategy of only reconstructing nontrivially when at most two packets are missing. Instead, we correct for missing packets and compute the probabilities for the residual mean-square error.

Lemma 4. *Let $\{\beta_1, \beta_2, \dots, \beta_n\}$, $n \geq 1$, be independent, identically distributed binary random variables with probability $\mathbb{P}(\beta_1 = 1) = q$. Let the random variables $\gamma_1, \gamma_2, \dots, \gamma_n$ be defined by $\gamma_j = \beta_j$ if $\sum_{j=1}^n \beta_j \geq 3$, and otherwise $\gamma_j = 0$ for all $j \in \{1, 2, \dots n\}$. Then, for any j,*

$$\mathbb{P}(\gamma_j = 1) \leq \frac{1}{6} n^3 q^4,$$

and for $j_1 \neq j_2$, we have

$$\mathbb{P}(\gamma_{j_1} = \gamma_{j_2} = 1) \leq n^2 q^4.$$

These lemmata allow us to formulate an error bound for the remaining mean-square error for blind reconstruction after the error correction protocol has been applied.

Theorem 1. *Let $V = V_1 \otimes V_2 \otimes \cdots \otimes V_m$ be the analysis operator of a Parseval product frame $\mathcal{F} = \mathcal{F}^{(1)} \otimes \mathcal{F}^{(2)} \otimes \cdots \otimes \mathcal{F}^{(m)}$ for a Hilbert space $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \cdots \otimes \mathcal{H}_m$. Denote the dimension of each $\mathcal{H}_i$ by d_i and the number of frame vectors in $\mathcal{F}^{(i)}$ by n_i. Let $\{\beta_{j_1,j_2,\dots j_m}\}$ be an m-parameter family of binary independent, identically distributed random variables, define $\{\beta^{(m-1)}_j\}$ as above, and let $\gamma_{j_1,j_2,\dots j_m} = \beta^{(m-1)}_{j_m}$ if $\sum_{j_m=1}^{n_m} \beta^{(m-1)}_{j_m} \geq 3$ and $\gamma_{j_1,j_2,\dots j_m} = 0$ otherwise, then*

$$\sigma^2(V, \gamma) \leq \frac{1}{d_m} \Big((q_m - r_m) \sum_{j=1}^{n_m} \|f_j^{(m)}\|^4 + r_m \sum_{j,l=1}^{n_m} |\langle f_j^{(m)}, f_l^{(m)} \rangle|^2 \Big)$$

with

$$q_m = 6^{1-2\cdot 3^{m-1}} n_m^3 n_{m-1}^{4\cdot 3^1} n_{m-2}^{4\cdot 3^2} \cdots n_1^{4\cdot 3^{m-1}} q_0^{4\cdot 3^{m-1}}$$

and

$$r_m \leq \frac{6}{n_m} q_m \,.$$

Corollary 1. *If $V = V_1 \otimes V_2 \otimes \cdots \otimes V_m$ and all V_i belong to equal-norm Parseval frames, then it is well known that $\|f_j^{(i)}\|^2 = \frac{d_i}{n_i}$ and by the Cauchy Schwarz inequality $|\langle f_j^{(i)}, f_l^{(i)} \rangle|^2 \leq d_i^2/n_i^2$. Thus, we have*

$$\sigma^2(V, \gamma) \leq q_m \frac{d_m}{n_m} + r_m d_m \leq 7 q_m \frac{d_m}{n_m}$$

with

$$q_m = 6^{1-2\cdot 3^{m-1}} n_m^3 n_{m-1}^{4\cdot 3^1} n_{m-2}^{4\cdot 3^2} \cdots n_1^{4\cdot 3^{m-1}} q_0^{4\cdot 3^{m-1}} \,.$$

Example 1. *Assume that an equal-norm product frame $\mathcal{F} = \mathcal{F}^{(1)} \otimes \cdots \otimes \mathcal{F}^{(m)}$ has $\mathcal{F}^{(i)}$ with $n_i = i^2 n_1$ vectors for each $i \in \{1, 2, \dots m\}$ and $n_1 \geq 3$. Let the dimension of the Hilbert space $\mathcal{H}_i$ spanned by $\mathcal{F}^{(i)}$ be*

$$\dim(\mathcal{H}_i) = i^2 n_1 - 2 \,,$$

and assume the frame can correct any two erased coefficients. Examples of such frames are the harmonic ones, see e.g. [2].

The tensor product of these m Hilbert spaces, $\mathcal{H} = \otimes_{i=1}^{m} \mathcal{H}_i$, has dimension

$$\dim(\mathcal{H}) = (m!)^2 n_1^m \prod_{i=1}^{m} \Big(1 - \frac{2}{i^2 n_1}\Big) .$$

This means, the coding rate R is bounded, independently of m, by

$$\begin{aligned} R &> \prod_{i=1}^{m} \Big(1 - \frac{2}{i^2 n_1}\Big) > (1 - \frac{2}{n_1}) \Big(1 - \frac{2}{n_1} \sum_{i=2}^{\infty} \frac{1}{i^2}\Big) \\ &= (1 - \frac{2}{n_1}) \Big(1 - \frac{2}{6n_1} \Big(\frac{\pi^2}{6} - 1\Big)\Big) . \end{aligned}$$

It is straightforward to check that $n_1 \geq 3$ ensures $R > 0$. The preceding theorem then states that after correcting erasures, the probability of an uncorrected block at the final level is

$$q_m \leq m^6 n_1^3 6^{1-2\cdot 3^{m-1}} q_0^{4\cdot 3^{m-1}} e^{4 \sum_{k=1}^{m-1} 3^{m-k} \ln(k^2 n_1)}$$

and upon estimating the sum in the exponent with Jensen's inequality,

$$2 \sum_{k=1}^{m-1} 3^{-k} \ln k \leq 2 \sum_{k=1}^{\infty} 3^{-k} \ln k \leq \ln \frac{3}{2} ,$$

we have

$$q_m \leq m^6 n_1^3 6^{1-2\cdot 3^{m-1}} q_0^{4\cdot 3^{m-1}} e^{2(3^m - 1) \ln n_1} e^{4\cdot 3^m \ln \frac{3}{2}} .$$

To achieve exponential decay of q_m in 3^m requires

$$-2 \ln 6 + 4 \ln q_0 + 6 \ln n_1 + 12 \ln \frac{3}{2} < 0 ,$$

which amounts to

$$\frac{27}{8\sqrt{6}} q_0 n_1^{3/2} < 1 .$$

Since $n_1 = 3$ is the smallest dimension to start the iteration, fast decay of the mean-square error needs $q_0 < 8\sqrt{2}/81 \approx 0.14$.
The number of transmitted frame coefficients is $(m!)^2 n_1^m$, so by Stirling's approximation $O(e^{(m+\frac{1}{2}) \ln m + m \ln n_1})$, whereas by the preceding corollary the decay of the mean-square error is of order $O(e^{-c3^m})$, for a suitable $c > 0$. This implies that the mean-square error decays faster than any inverse power of the number of transmitted coefficients.

Acknowledgment

This work was partially supported by National Science Foundation grant DMS 08-07399 and by the Deutsche Forschungsgemeinschaft under Heisenberg Fellowship KU 1446/8-1.

References:

[1] V. K. Goyal, M. Vetterli, and N. T. Thao, Quantized overcomplete expansions in R^n: analysis, synthesis, and algorithms. *IEEE Trans. Inform. Theory*, 44(1): 16–31, 1998.

[2] V. K. Goyal, J. Kovačević, and J. A. Kelner, "Quantized frame expansions with erasures," *Appl. Comp. Harm. Anal.*, 10:203–233, 2001.

[3] J. Kovačević, P. L. Dragotti, and V. K. Goyal, "Filter bank frame expansions with erasures," *IEEE Trans. Inform. Theory*, 48:1439–1450, 2002.

[4] P. Casazza and J. Kovačević, "Equal-norm tight frames with erasures," *Adv. Comp. Math.*, 18:387–430, 2003.

[5] G. Rath and C. Guillemot, Performance analysis and recursive syndrome decoding of DFT codes for bursty erasure recovery, *IEEE Trans. on Signal Processing*, 51 (5):1335–1350, 2003.

[6] G. Rath and C. Guillemot, Frame-theoretic analysis of DFT codes with erasures, *IEEE Transactions on Signal Processing*, 52 (2):447–460, 2004.

[7] R. Holmes and V. I. Paulsen, "Optimal frames for erasures," *Lin. Alg. Appl.*, 377:31–51, 2004.

[8] B. G. Bodmann and V. I. Paulsen, "Frames, graphs and erasures," *Linear Algebra Appl.*, 404: 118–146, 2005.

[9] D. Kalra, Complex equiangular cyclic frames and erasures, Linear Algebra Appl., 419:373–399, 2006.

[10] M. Püschel and J. Kovačević, "Real, tight frames with maximal robustness to erasures", Proc. Data Compr. Conf., Snowbird, UT, 63–72, March 2005.

[11] P. G. Casazza and G. Kutyniok, Robustness of fusion frames under erasures of subspaces and of local frame vectors, Contemp. Math., 464, Amer. Math. Soc., Providence, RI, 149–160, 2008.

[12] P. G. Casazza, G. Kutyniok, and S. Li, "Fusion Frames and Distributed Processing," *Appl. Comput. Harmon. Anal.*, 25:114–132, 2008.

[13] G. Kutyniok, A. Pezeshki, A. R. Calderbank, and T. Liu, "Robust Dimension Reduction, Fusion Frames, and Grassmannian Packings," *Appl. Comput. Harmon. Anal.*, 26:64–76, 2009.

[14] P. G. Casazza and G. Kutyniok, "Frames of subspaces," in: "Wavelets, frames and operator theory," Contemp. Math., 345, Amer. Math. Soc., Providence, RI, 87–113, 2004.

[15] B. G. Bodmann, "Optimal linear transmission by loss-insensitive packet encoding," *Appl. Comput. Harmon. Anal.*, 22:274–285, 2007.

[16] P. Elias, Error-free coding, *IRE Trans. IT*, 4:29–37, 1954.

[17] O. Christensen, "An Introduction to Frames and Riesz Bases," Birkhäuser, Boston, 2003.

Representation of operators by sampling in the time-frequency domain

Monika Dörfler [(1)] and Bruno Torrésani[(2)]

(1) ARI, Austrian Academy of Science, Wohllebengasse 12-14, A-1040 Vienna, Austria.
(2) LATP, Centre de Mathématique et d'Informatique, 39 rue Joliot-Curie, 13453 Marseille cedex 13, France.
Monika.Doerfler@oeaw.ac.at, Bruno.Torresani@cmi.univ-mrs.fr

Abstract:

Gabor multipliers are well-suited for the approximation of certain time-variant systems. However, this class of systems is rather restricted. To overcome this restriction, multiple Gabor multipliers allowing for more than one synthesis windows are introduced. The influence of the choice of the various parameters involved on approximation quality is studied for both classical and multiple Gabor multipliers.

1. Introduction

In a recent paper [1], the authors describe the representation of operators in the time-frequency domain by means of a twisted convolution with the operator's spreading function. Although not suitable for direct discretization, the spreading representation provides a better understanding of certain operators' behavior: it reflects the operator's action in the time-frequency domain. This motivates an approach that uses the spreading representation of time-frequency multipliers [1], in order to optimize the parameters involved. More specifically, in the one-dimensional, continuous-time case, given an operator H with integral kernel κ_H and spreading function η_H:

$$\eta_H(b,\nu) = \int_{-\infty}^{\infty} \kappa_H(t,t-b)e^{-2i\pi\nu t}\,dt,$$

we aim at modeling the operator by its action on the sampled short-time Fourier transform (STFT) or Gabor coefficients, given for any $f \in \mathbf{L}^2(\mathbb{R})$ by

$$\mathscr{V}_g f(mb_0, n\nu_0) = \langle f, g_{mn}\rangle\,, \quad m,n \in \mathbb{Z} \qquad (1)$$

where the $g_{mn} = M_{n\nu_0}T_{mb_0}g$ denote the Gabor atoms associated to $g \in \mathbf{L}^2(\mathbb{R})$ and the lattice constants $b_0, \nu_0 \in \mathbb{R}^+$, see [3][1]. In the case of classical Gabor multipliers, the modification consists of a pure multiplication. Thus, the linear operator applied to the coefficients $\mathscr{V}_g f$ is diagonal, an approach that leads to accurate approximation for so-called underspread operators [5]. The restriction to diagonality may be relaxed in order to achieve better approximation for a wider class of operators at low cost. It also appears, that in certain approximation tasks it is more efficient, e.g. in the sense of sparsity, to use several side diagonals, but a lower redundancy in the Gabor system used.

The aim of this contribution is the description of error estimates for the approximation of operators by generalized Gabor multipliers, based on the operator's spreading function. From this description guidelines for the choice of good parameters for the approximation are deduced and illustrated by various numerical experiments.

[1]The finite dimensional case $\mathcal{H} = \mathbb{C}^L$ is obtained similarly, replacing integrals with finite sums, and letting $m = 0,\dots N_b - 1$, $n = 0,\dots N_\nu - 1$, where $N_b = L/b_0$, $N_\nu = L/\nu_0$ and b_0, ν_0 divide L.

2. Approximation in the time-frequency domain: the parameters

Throughout this paper, $\mathcal{H}$ denotes a (finite or infinite-dimensional) Hilbert space, equipped with an action of the Heisenberg group of time-frequency shifts.

2.1 Time-frequency multipliers

Let $\mathscr{V}_g^*$ denote the adjoint of $\mathscr{V}_g$. A Gabor multiplier [4] is defined as

$$\mathbb{M} : f \in \mathcal{H} \longmapsto \mathbb{M}f = \mathscr{V}_2^*(\mathbf{m}\cdot\mathscr{V}_1 f).$$

Here, $\mathbf{m}$ is the pointwise multiplication operator whose symbol, defined on the lattice Λ will also be denoted by $\mathbf{m}$. We shall denote by Λ^o the adjoint lattice, $\square^o$ its fundamental domain, and Π^o the corresponding periodization operator. In the infinite-dimensional situation $\mathcal{H} = \mathbf{L}^2(\mathbb{R})$, and for a product lattice of the form $\Lambda = b_0\mathbb{Z}\times\nu_0\mathbb{Z}$, we have $\Lambda^o = t_0\mathbb{Z}\times\xi_0\mathbb{Z}$ with $t_0 = 1/\nu_0$, $\xi_0 = 1/b_0$, and $\Pi^o f(\zeta) = \sum_{\lambda^o\in\Lambda^o} f(\zeta+\lambda^o)$, $\zeta \in \square^o$. In a finite-dimensional setting $\mathcal{H} = \mathbb{C}^L$, with $\Lambda = \mathbb{Z}_{N_b}\times\mathbb{Z}_{N_\nu}$, with N_b, N_ν two divisors of L, we have $\Lambda^o = \mathbb{Z}_{N_\nu}\times\mathbb{Z}_{N_b}$, and the obvious form for the periodization operator.

In the definition of the multipliers, several parameters have to be fixed: the analysis and synthesis windows g and h, the lattice Λ, and the symbol $\mathbf{m}$. For practical as well as theoretical reasons, the windows should be well-localized in time and frequency. As for the lattice, it is expected that denser lattices will lead to better results in approximation, but higher computational cost. However, it will be seen that too dense lattices are not suitable.

Finally, the symbol $\mathbf{m}$ can be optimized to best approximate a given operator. In [1], an explicit expression for the best approximation was obtainned in the spreading domain, yielding a very efficient algorithm (compare [2]).

The spreading function of Gabor multipliers takes the form $\eta_{\mathbb{M}}(\zeta) = \mathscr{M}(\zeta) \cdot \mathscr{V}_g h(\zeta)$, where $\mathscr{M}$ is the symplectic Fourier transform of $\mathbf{m}$. Note, that this leads to a periodic function with period $\square^o$. Hence, good approximation by a classical Gabor multiplier is possible, if the essential support of the spreading function is smaller than 1 and can then be contained in the fundamental domain $\square^o$ of the adjoint lattice for a dense enough lattice Λ. Also, to reduce aliasing as much as possible, the analysis and synthesis windows must be chosen such that $\mathscr{V}_g h$ is small outside $\square^o$ and positive on the support of the spreading function, also see Section 4.1.

2.2 Generalized Gabor multipliers

Multiple Gabor multipliers are sums of Gabor multipliers with different synthesis windows.

Definition 1 (Multiple Gabor Multiplier) *Let $g, h \in \mathcal{H}$ denote two window functions. Let Λ be a time-frequency lattice. Let $\{\mu_j, j \in J\}$ denote a finite set of time-frequency shifts, and let $\{\mathbf{m}_j, j \in J\}$ be a family of bounded functions on Λ. Set $h^{(j)} = \pi(\mu_j)h$, then the associated generalized Gabor multiplier $\mathbb{M}$ is defined, for $f \in \mathcal{H}$, as*

$$\mathbb{M}f = \sum_{\lambda\in\Lambda} \sum_{j\in J} \mathbf{m}(\lambda, \mu_j)\langle f, \pi(\lambda)g\rangle \pi(\lambda)h^{(j)} .$$

It is immediately obvious that in addition to the parameters mentioned above, the window h as well as the sampling points J must be chosen.

3. Error analysis in $\mathbf{L}^2(\mathbb{R})$

In [1], it was shown that the symbol $\mathbf{m}(\lambda, \mu_j) := \mathbf{m}_j(\lambda)$ of the best approximation of a Hilbert-Schmidt operator by a multiple Gabor multiplier with fixed sets Λ, J and windows, is given by the symplectic Fourier transform of the $\square^o$-periodic functions $\mathscr{M}_j$ obtained via the vector equation

$$\mathscr{M}(\zeta) = \mathcal{U}(\zeta)^{-1} \cdot \mathcal{B}(\zeta) , \quad \zeta \in \square^o , \tag{2}$$

where the matrix and vector valued functions $\mathcal{U}$ and $\mathcal{B}$ are given by the Λ^o-periodizations

$$\mathcal{U}_{jj'} = \Pi^o\Big(\mathscr{V}_g h^{(j')}\overline{\mathscr{V}_g h^{(j)}}\Big) \ , \ \mathcal{B}_j = \Pi^o\Big(\eta_H \overline{\mathscr{V}_g h^{(j)}}\Big) \ ,$$

provided $\mathcal{U}$ is invertible a.e.

The case of one synthesis windows may be immediately obtained from the above formula. Note that formula (2) allows for an efficient implementation of the otherwise expensive calculation of the best approximation by multiple Gabor multipliers.

We may now give an expression for the error in the approximation given above, in the case $\mathcal{H} = \mathbf{L}^2(\mathbb{R})$

Proposition 1 *Let $\mathscr{M}$ denote the vector-valued function obtained as in* (2) *and set, for the Hilbert-Schmidt operator H, $\Gamma_H = \Pi^o(|\eta_H|^2)$. Then the approximation error $E = \|\eta_H - \sum_j \mathscr{M}_j \mathscr{V}_j\|^2$ is given by*

$$E = \int_{\square^o} |\Gamma_H(\zeta)| \Big(1 - \frac{\sum_{i,j}(\mathcal{U}^{-1})_{ij}(\zeta)\mathcal{B}_i(\zeta)\overline{\mathcal{B}_j}(\zeta)}{|\Gamma_H(\zeta)|}\Big) d\zeta$$

Notice that this covers the multiplier case obtained in [1]. Notice also that this immediately yields

$$E \le \|\eta_H\|^2 \left\| 1 - \frac{\sum_{i,j}(\mathcal{U}^{-1})_{ij}\mathcal{B}_i\overline{\mathcal{B}_j}}{|\Gamma_H|} \right\|_\infty$$

The finite-dimensional situation is similar, replacing the integral over $\square^o$ with a finite sum over the finite fundamental domain $\{0, \dots t_0 - 1\} \times \{0, \dots \xi_0 - 1\}$.

4. Choosing the parameters

For simlicity, we specialize the following discussion to the infinite-dimensional case $\mathcal{H} = \mathbf{L}^2(\mathbb{R})$, and rectangular lattice $\Lambda = b_0\mathbb{Z} \times \nu_0\mathbb{Z}$. The finite-dimensional situation is handled similarly.

4.1 Gabor Multipliers

If an operator with known spreading function is to be approximated by a Gabor multiplier, the lattice may be adapted to the eccentricity of the spreading function according to the error expression obtained in Proposition 1, which may be considerably simplified for the case of only one synthesis window, see [1]. In order to choose the eccentricity of the lattice accordingly and adapt the window to the chosen lattice as to avoid aliasing, assume, that we may find b_0, ν_0, with $b_0 \cdot \nu_0 < 1$, such that $\mathrm{supp}(\eta_H) \subseteq T_z\square^o$, where $\square^o = [0, \frac{1}{\nu_0}] \times [0, \frac{1}{b_0}]$. In this case, the error resulting from best approximation by a Gabor multiplier with respect to the lattice $b_0\mathbb{Z} \times \nu_0\mathbb{Z}$ is bounded by $C_e \cdot \|\eta_H\|_2^2$, with

$$C_e = 1 - \inf_{t,\xi\in\square_H^o} \frac{|\mathscr{V}_g h(t,\xi)|^2}{\sum_{k,l} |\mathscr{V}_g h(t + kt_0, \xi + l\xi_0)|^2}, \tag{3}$$

with $\square_H^o = \square^o \cap \mathrm{Supp}(\eta_H)$, and becomes minimal for a window that is optimally concentrated inside $\square^o$. Heuristically as well as from numerical experiments we know, that the tight window, [3], corresponding to the given lattice is usually a good choice to fulfill this requirement.

4.2 Generalized Gabor Multipliers

The main additional task in the generalized situation is the choice of the sampling points μ_j for the synthesis windows. A good choice will again be guided by the behavior of the spreading function. The relevant areas in the spreading domain should be covered as well as possible with the smallest possible overlap by the cross-ambiguity functions of the different synthesis windows with respect to a given reference-window localized at $(0, 0)$ e.g. the Gaussian window. Motivated by the results from the Gabor multiplier situation, we choose a tight window with respect to the analysis lattice and look for the most appropriate sampling points for the synthesis windows. Examples will be given in Section 5.2.

5. Examples

We now turn to numerical experiments, in the finite case $\mathcal{H} = \mathbb{C}^L$. In the following examples, the relative approximation error for the best approximation $\tilde{H}$ of H is given

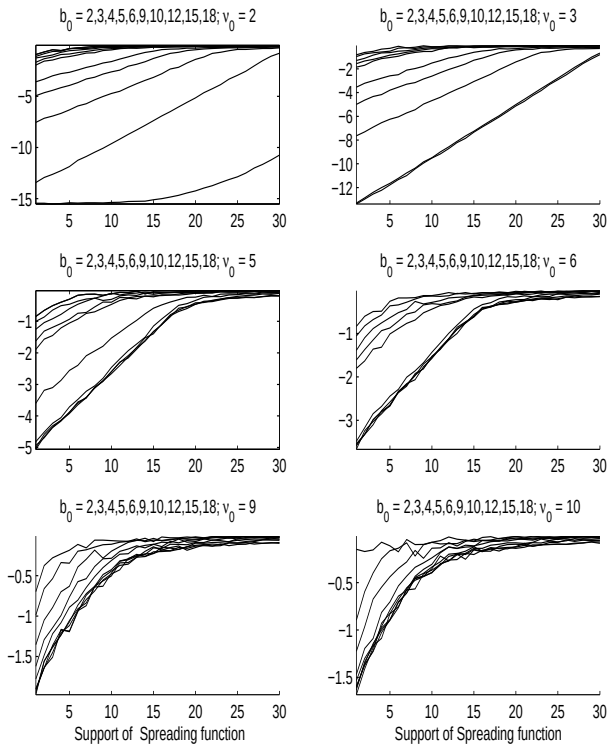


Figure 1: *Approximation error for different bandwidth of spreading function and different values of* b_0, ν_0.

by

$$E = \|\tilde{H} - H\|/\|H\| \,,$$

the logarithm of which is represented in the next plots. We display here the Fröbenius norm, the plots obtained with the operator norm are almost identical.

5.1 Classical Gabor Multipliers

We generate operators with compact support in the spreading domain, in a square of side size between 3 and 61, symmetric about 0. The values are random, the signal length is $L = 180$. We then investigate the approximation quality for various pairs of lattice constants, with b_0 varying between 2 and 18 and ν_0 between 2 and 10. The results are presented in Figure 1. Note the two distinct regimes: the error grows exponentially up to a certain value of the support size, depending on the lattice density, and slower thereafter. A possible explanation for this effect, to be further investigated, is the fact, that the error (see the bound in (3)) is comprised of an aliasing error and the inherent inaccuracy of Gabor multiplier approximation, even for very high sampling density, of overspread operators.

In order to emphasize the importance of lattice adaptation to eccentricity, we show the results for different lattice constants resulting in the same redundancy (5) in Figure 2. The solid lines show the results for $b_0 = \nu_0 = 6$, leading to far better results than the lattice constants not adapted to the (symmetric) support of the spreading function.

5.2 Generalized Gabor Multipliers

In order to illustrate the influence of additional synthesis windows on the approximation quality, we first consider

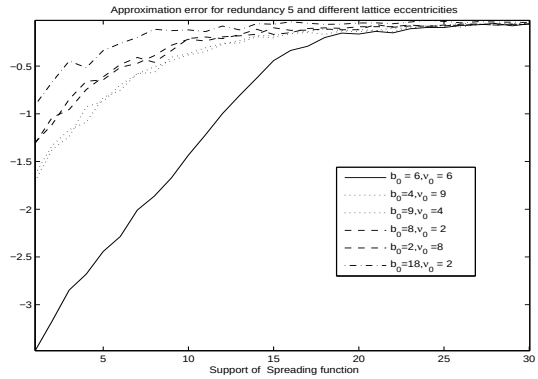


Figure 2: *Approximation error for different lattice-eccentricity*

.

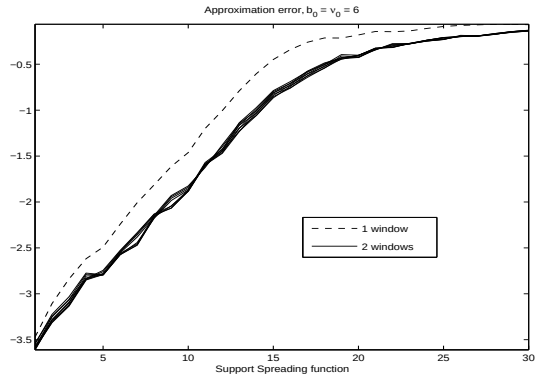


Figure 3: *Spreading function of operator and best approximation with one or two synthesis windows, approximation error for growing support of spreading function.*

the same operators as in the previous section, but allow for one additional synthesis window. Here, and in the subsequent examples, one window will always be a window centered about 0, as above, with a time-shifted version of the original window as additional window. Hence, only the shift-parameter of the additional window has to be considered. Figure 3 shows the improvement in approximation quality for shift-parameters of the additional window between -5 and 5 (solid), as opposed to the single window approximation.

Next, we investigate the following situation: an operator with two effectively disjoint components in the spreading domain is, again, approximated by a multiple Gabor multiplier with 2 synthesis windows. For better comparison, the two components are the component from the previous examples plus a shifted version (by 90 samples) thereof. Figure 4 shows the spreading functions of one of the operators and its best approximation with two synthesis windows, for the optimal additional window. Note the aliasing effect. In this situation, using two appropriate synthesis windows, the obtained results are similar to those in the case of one spreading function component and one synthesis window, as discussed in the previous section. In Figure 5, we display the results for 3 symmetric pairs of lattice constants, the optimal window's result being represented by the solid line, while the dashed lines show the results of close but suboptimal synthesis windows. As the operator was generated by a translation by 90 samples, the

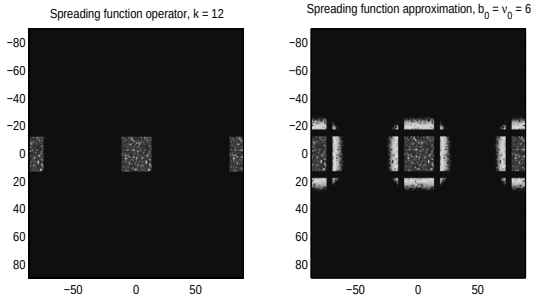


Figure 4: *Spreading function of operator and best approximation.*

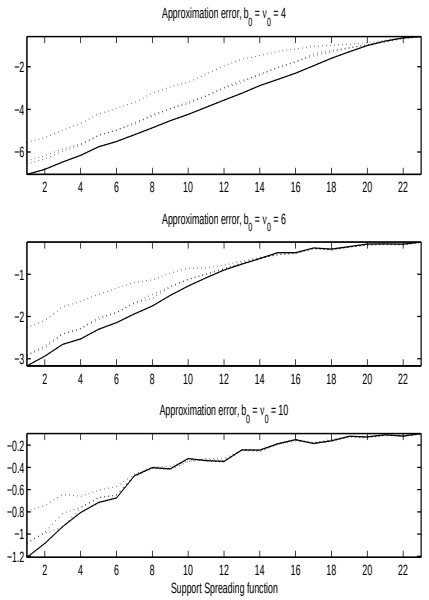


Figure 5: *Approximation error for varying support of two components of spreading function and two synthesis windows.*

tight window, shifted by 90 samples itself, is expected to be the optimal additional window. This is confirmed by the experiments.

In a last experiment, the two components in the spreading domain are close and, for growing bandwidth, overlapping. Figure 6 shows, as before, the results of approximation for growing support of both spreading function components, with $b_0 = \nu_0 = 6$ and various additional synthesis windows. The additional window with shift-parameter 0 is, of course, the original window and yields the approximation result obtained for a single synthesis window. For the optimal window, the result is close to the single window/single component case for the same lattice.

6. Discussion and conclusions

The examples given in the previous section show that the choice of various parameters has considerable influence on the performance of approximation by (generalized) Gabor multipliers. While the situation is rather easily understood in the case of classical Gabor multipliers, it is much more intricate in the generalized case. It should be noted that, while yielding better results in the approximation, using a small number of additional synthesis windows does not dramatically increase the computational cost: in (2),

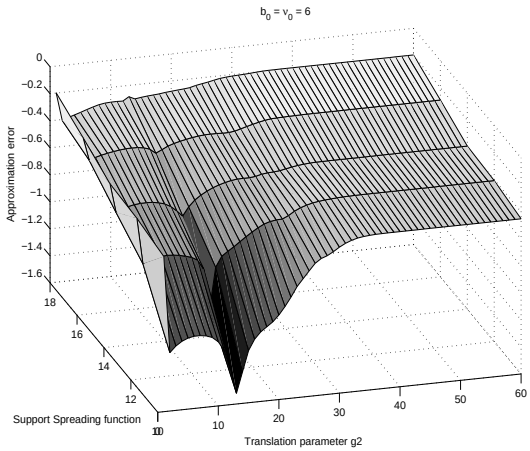


Figure 6: *Approximation error for growing support of spreading function and various additional synthesis windows.*

going from $|J| = 1$ to larger index sets J involves inverting (generally small) matrices instead of computing a point-wise ratio. Higher redundancy of the Gabor system involved is more expensive in the sense of coefficients. In many cases, using an additional window may be more favorable in improving approximation quality than a denser lattice. Future work on this topic will include systematic numerical experiments as well as the analytical investigation of the approximation quality of generalized and classical Gabor multipliers. Another goal is the development of a method to determine an adapted sampling scheme for the synthesis windows from an operator's spreading function.

7. Acknowledgments

The first author was funded by project MA07-025 of WWTF Austria. The second author was partly supported by the CNRS programme PEPS/ST2I *MTF&Sons*.

References:

[1] Monika Dörfler and Bruno Torrésani. Representation of operators in the time-frequency domain and generalized Gabor multipliers. *arXiv:0809.2698*, 2008, to appear in *Journal of Fourier Anal. and Appl.*

[2] Hans G. Feichtinger, Mario Hampejs, and Günther Kracher. Approximation of matrices by Gabor multipliers. IEEE Signal Proc. Letters, 11(11):883– 886, 2004.

[3] Hans G. Feichtinger and Thomas Strohmer. *Gabor Analysis and Algorithms. Theory and Applications.* Birkhäuser, 1998.

[4] Hans Georg Feichtinger and Kristof Nowak. A first survey of Gabor multipliers. In H. G. Feichtinger and T. Strohmer, editors, *Advances in Gabor Analysis*, Boston, 2002. Birkhauser.

[5] Werner Kozek. Adaptation of Weyl-Heisenberg frames to underspread environments. In [3], 1998.

Operator Identification and Sampling

Götz Pfander [1] and David Walnut [2]

(1) School of Engineering and Science, Jacobs University Bremen, 28759 Bremen, Germany.
(2) Dept. of Mathematical Sciences, George Mason University, Fairfax, VA 22030 USA.
g.pfander@iu-bremen.de dwalnut@gmu.edu

Abstract:

Time–invariant communication channels are usually modelled as convolution with a fixed impulse–response function. As the name suggests, such a channel is completely determined by its action on a unit impulse. Time–varying communication channels are modelled as pseudodifferential operators or superpositions of time and frequency shifts. The function or distribution weighting those time and frequency shifts is referred to as the spreading function of the operator. We consider the question of whether such operators are identifiable, that is, whether they are completely determined by their action on a single function or distribution. It turns out that the answer is dependent on the size of the support of the spreading function, and that when the operators are identifiable, the input can be chosen as a distribution supported on an appropriately chosen grid. These results provide a sampling theory for operators that can be thought of as a generalization of the classical sampling formula for bandlimited functions.

1. Channel Models and Identification

A communications channel is said to be *measurable* or *identifiable* if its characteristics can be determined by its action on a single fixed input signal. A general model for linear (time-varying) communication channels is as operators of the form

$$Hf(x) = \int h_H(t,x)\, f(x-t)\, dt.$$

The function $h_H(t,x)$ is referred to as the *impulse response* of the channel and is interpreted as the response of the channel at time x to a unit impulse at time $x-t$, that is, originating t time units earlier. If $h_H(t,x) = h_H(t)$ then the characteristics of the channel are time-invariant and in this case the channel is modelled as a convolution operator. Such channels are identifiable since $h_H(t)$ can be recovered as the response of the channel to the input signal $\delta_0(t)$, the unit-impulse at $t=0$.

There are two representations of H that will be convenient for our purposes.

1. Letting $\eta_H(t,\nu) = \int h_H(t,x)\, e^{-2\pi i\nu(x-t)}\, dx$ gives

$$\begin{aligned} Hf(x) &= \iint \eta_H(t,\nu)\, e^{2\pi i\nu(x-t)}\, f(x-t)\, d\nu\, dt \\ &= \iint \eta_H(t,\nu)\, T_t\, M_\nu f(x)\, d\nu\, dt. \end{aligned}$$

$\eta_H(t,\nu)$ is the *spreading function* of H. If $\operatorname{supp}\eta_H \subseteq [0,a]\times[-b/2,b/2]$ for some $a,\, b>0$ then a is called the *maximum time-delay* and b the *maximum Doppler spread* of the channel.

2. Letting $\sigma_H(x,\xi) = \int h_H(t,x)\, e^{2\pi i t\xi}\, dt$ gives

$$Hf(x) = \int \sigma_H(x,\xi)\widehat{f}(\xi)\, e^{2\pi i x\xi}\, d\xi.$$

$\sigma_H(x,\xi)$ is the *Kohn-Nirenberg* (KN) symbol of H and we have the relation

$$\eta_H(t,\nu) = \iint \sigma_H(x,\xi)\, e^{-2\pi i(\nu x-\xi t)}\, dx\, d\xi.$$

In other words, the spreading function η_H is the *symplectic Fourier transform* of the KN symbol of H.

In 1963, T. Kailath [3, 4, 5] asserted that for time-variant communication channels to be identifiable it is necessary and sufficient that the maximum time-delay, a, and Doppler shift, b, satisfy $ab \le 1$ and gave an argument for this assertion based on counting degrees of freedom. In the argument, Kailath looks at the response of the channel to a train of impulses separated by at least a time units, so that in this sense the channel is being "sampled" by a succession of evenly-spaced impulse responses. The condition $ab \le 1$ allows for the recovery of sufficiently many samples of $h_H(t,x)$ to determine it uniquely.

Kailath's conjecture was given a precise mathematical framework and proved in [6]. The framework is as follows. Choose normed linear spaces $D(\mathbf{R})$ and $Y(\mathbf{R})$ of functions or distributions on $\mathbf{R}$, and a normed linear space of bounded linear operators $\mathcal{H} \subset \mathcal{L}(D(\mathbf{R}), Y(\mathbf{R}))$. Each fixed element $g \in D(\mathbf{R})$ induces a map $\Phi_g : \mathcal{H} \longrightarrow Y(\mathbf{R}),\ H \mapsto Hg$. If for some $g \in D(\mathbf{R})$, Φ_g is bounded above and below, that is, there are constants $0 < A \le B$ such that for all $H \in \mathcal{H}$,

$$A\|H\|_{\mathcal{H}} \le \|Hg\|_Y \le B\,\|H\|_{\mathcal{H}}$$

then we say that $\mathcal{H}$ is *identifiable with identifier* $g \in D(\mathbf{R})$.

Taking $D = S_0'$, $Y = L^2$, and $\mathcal{H}_S = \{H \in HS(L^2)\colon \eta_H \in S_0(\mathbf{R} \times \widehat{\mathbf{R}}),\ \operatorname{supp}\eta_H \subseteq S\}$ where $S \subseteq \mathbf{R} \times \widehat{\mathbf{R}}$, $HS(L^2)$ is the class of Hilbert-Schmidt operators, and S_0 is the Feichtinger algebra (defined below), the following was proved in [6].

Theorem 1. If $S = [0, a] \times [-b/2, b/2]$ then $\mathcal{H}_S$ is identifiable if and only if $ab \leq 1$. In this case an identifier is given by $g = \sum_n \delta_{na}$.

2. Distributional Spreading Functions and Operator Sampling

The requirement that $\eta_H \in S_0$ excludes some very natural operators from consideration in this formalism, for example the identity operator ($\eta_H(t, \nu) = \delta_0(t)\delta_0(\nu)$), convolution operators ($\eta_H(t, \nu) = h(t)\delta_0(\nu)$ giving $Hf = f * h$), and multiplication operators, ($\eta_H(t, \nu) = \delta_0(t)\widehat{m}(\nu)$ giving $Hf = m \cdot f$).

A more natural setting for operator identification is the *modulation spaces* (see [2] for a full treatment of the subject). For convenience we give the definitions below for modulation spaces on $\mathbf{R}$, but all definitions and results can be extended to $\mathbf{R}^d$. For $\varphi \in \mathcal{S}(\mathbf{R})$ define for $f \in \mathcal{S}'(\mathbf{R})$ the *short-time Fourier transform (STFT)* of f by

$$\begin{aligned} V_\varphi f(t, \nu) &= \langle f, T_t M_\nu \varphi \rangle \\ &= \int f(x)\, e^{-2\pi i \nu(x-t)}\, \overline{\varphi(x-t)}\, dx. \end{aligned}$$

For $1 \leq p, q \leq \infty$ define the modulation space $M^{p,q}(\mathbf{R})$ by

$$M^{p,q}(\mathbf{R}) = \{f \in \mathcal{S}'(\mathbf{R})\colon V_\varphi f \in L^{p,q}(\mathbf{R})\},$$

that is, for which

$$\|V_\varphi\|_{L^{p,q}} = \left(\int \left(\int |V_\varphi f(t, \nu)|^p\, dt\right)^{q/p}\right)^{1/q}$$

is finite. The usual modifications are made if p or $q = \infty$. $M^{p,q}$ is a Banach space with respect to the norm $\|f\|_{M^{p,q}} = \|V_\varphi f\|_{L^{p,q}}$ and different nonzero choices of $\varphi \in \mathcal{S}$ define equivalent norms. The space $M^{1,1}$ is the Feichtinger algebra denoted S_0 and $M^{\infty,\infty}$ is its dual S_0'. The space S_0' contains the Dirac impulses $\delta_x : f \mapsto f(x)$ for $x \in \mathbf{R}$ as well as distributions of the form $g = \sum_j c_j \delta_{x_j}$, $x_j \in \mathbf{R}$ and $\{c_j\} \subseteq \mathbf{C}$ a bounded sequence.

In our next step toward operator sampling we observe that it is possible to take $D = S_0'$, $Y = S_0'$, and $\mathcal{H}_S = \{H \in \mathcal{L}(D, Y)\colon \eta_H \in S_0',\ \operatorname{supp}\eta_H \subseteq S\}$ in the operator identification formalism. Indeed the following theorem was shown in [10].

Theorem 2. The operator class $\mathcal{H}_S$ (defined above) is identifiable if $S = [0, a] \times [-b/2, b/2]$ and $ab < 1$, and is not identifiable if $ab > 1$.

3. A Theory of Operator Sampling

In discussing identifiability of operators in various settings, we have been content to show that an operator is completely determined by its actions on a fixed input in terms of a norm inequality. The next step is to find an explicit reconstruction formula for the impulse response of the channel operator directly from its response to the identifier. Such formulas illustrate a connection between operator identification and classical sampling theory and lead to a definition of *operator sampling*.

If, in the operator identification formalism described earlier, an operator class $\mathcal{H}$ is identified by a distribution of the form $g = \sum_j c_j \delta_{x_j}$, then we call $\{x_j\}$ a *set of sampling* for $\mathcal{H}$ and g a *sampling function* for the operator class $\mathcal{H}$. In the results obtained so far, operator sampling is possible only for operators with compactly supported spreading function, and in order to interpret Theorem 1 in this context we make the following definition.

Given a Jordan domain $S \subseteq \mathbf{R}^2$, define the *operator Paley-Wiener space* $OPW^2(S)$ by

$$OPW^2(S) = \{H \in HS(L^2)\colon \operatorname{supp}\eta_H \subseteq S\}.$$

OPW^2 is a Banach space with respect to the Hilbert-Schmidt norm $\|H\|_{OPW^2} = \|\eta_H\|_{L^2}$. Then Theorem 1 can be extended as follows ([8]).

Theorem 3. Let Ω, T, $T' > 0$ with $T' < T$ and $\Omega T < 1$. Then $OPW^2([0, T'] \times [-\Omega/2, \Omega/2])$ is identifiable with identifier $g = \sum_n \delta_{nT}$ and moreover we have the formula

$$h_H(t, x) = r(t) \sum_{k \in \mathbf{Z}} (Hg)(t + kT)\varphi(x - t - kT)$$

unconditionally in $L^2(\mathbf{R}^2)$, where $r \in \mathcal{S}(\mathbf{R})$ is such that $r = 1$ on $[0, T']$ and vanishes outside a sufficiently small neighborhood of $[0, T']$, and where $\varphi \in \mathcal{S}(\mathbf{R})$ is such that $\widehat{\varphi} = 1$ on $[-\Omega/2, \Omega/2]$ and vanishes outside a sufficiently small neighborhood of $[-\Omega/2, \Omega/2]$.

In the more general modulation space setting we can define the operator Paley-Wiener space $OPW^{p,q}(S)$ by

$$\begin{aligned} OPW^{p,q}(S) &= \{H \in \mathcal{L}(S_0, S_0') \\ &\quad : \operatorname{supp}\eta_H \subseteq S,\ \ \sigma_H \in M^{pq,11}\} \end{aligned}$$

where $\sigma_H(x, \xi) \in M^{pq,11}$ means that the two-dimensional STFT of σ_H satisfies

$$\int \left(\int \left(\int |V_{\varphi\otimes\varphi}\sigma_H(t_1, t_2, \nu_1, \nu_2)|^p dt_1\right)^{q/p} dt_2\right)^{1/p} d\nu_1 d\nu_2$$

is finite. Here

$$V_{\varphi\otimes\varphi}(t_1, t_2, \nu_1, \nu_2) = \langle f, T_{t_1} M_{\nu_1}\varphi \otimes T_{t_2} M_{\nu_2}\varphi\rangle.$$

$OPW^{p,q}$ is a Banach space with respect to the norm $\|H\|_{OPW^{p,q}} = \|\sigma_H\|_{M^{pq,11}}$. In this case, Theorem 3 generalizes as follows ([8]).

Theorem 4. Let $1 \leq p, q \leq \infty$, Ω, T, $T' > 0$ with $T' < T$ and $\Omega T < 1$. Then $OPW^{p,q}([0, T'] \times [-\Omega/2, \Omega/2])$ is identifiable with identifier $g = \sum_n \delta_{nT}$ and moreover we have the formula

$$h_H(t, x) = r(t) \sum_{k \in \mathbf{Z}} (Hg)(t + kT)\varphi(x - t - kT)$$

unconditionally in $M^{1p,q1}(\mathbf{R}^2)$ and in the weak-* sense if p or $q = \infty$, where r and φ are as in Theorem 3.

Example 1. If we take H to be ordinary convolution by $h_H(t)$, this means that $h_H(t,x)$ depends only on t, that is, $h_H(t,x) = h_H(t)$. In this case H can be identified in principle by $g = \delta_0$, the unit impulse, since $Hg(x) = h_H(x)$. Translating this into our operator sampling formalism results in something slightly different.
Assume that $h \in M^{1,q}$ is supported in the interval $[0,T']$ and that $T > T'$, and $\Omega > 0$ are chosen so that $\Omega T < 1$. In this case, $\eta_H(t,\nu) = h(t)\,\delta_0(\nu)$ and $\sigma_H(x,\xi) = \widehat{h}(\xi)$. Therefore $\sigma_H \in M^{\infty q,11}$ and $H \in OPW^{\infty,q}([0,T'] \times \{0\})$.
If $g = \sum_n \delta_{nT}$ then Hg is simply the T–periodized impulse response $h(t)$, and it follows that

$$\begin{aligned} & r(t)\sum_{k\in\mathbf{Z}}(Hg)(t+kT)\varphi(x-t-kT) \\ = \;& r(t)\,h(t)\sum_{k\in\mathbf{Z}}\varphi(x-t-kT) \;=\; h(t) \end{aligned}$$

since $r(t) = 1$ on $[0,T']$ and vanishes outside a neighborhood of $[0,T']$ and since $\sum_k \varphi(x-t-kT) = 1$ by the Poisson Summation Formula and in consideration of the support constraints on $\widehat{\varphi}$. Indeed the theorem says that the sum $\sum_k \varphi(x-t-kT)$ converges to 1 in the $M^{\infty,1}$ norm and in particular uniformly on compact sets.

Example 2. If we take H to be multiplication by some fixed function $m \in M^{p,1}$ with $\operatorname{supp}\widehat{m} \subseteq [-\Omega/2, \Omega/2]$ then $\eta_H(t,\nu) = \delta_0(t)\widehat{m}(\nu)$, $h(t,x) = \delta_0(t)\,m(x-t)$, and $\sigma_H(x,\xi) = m(x)$. Therefore $\sigma_H \in M^{p\infty,11}$ and $H \in OPW^{p,\infty}(\{0\} \times [-\Omega/2, \Omega/2])$.
If $g = \sum_n \delta_{nT}$, with $T > 0$ chosen small enough that $\Omega T < 1$, then $Hg = \sum_n m(nT)\,\delta_{nT}$, and it follows from Theorem 4 that

$$\begin{aligned} & \delta_0(t)\,m(x-t) \\ = \;& r(t)\sum_{k\in\mathbf{Z}}(Hg)(t+kT)\varphi(x-t-kT) \\ = \;& r(t)\sum_{k\in\mathbf{Z}}\sum_{n\in\mathbf{Z}} m(nT)\,\delta_{(n-k)T}(t)\varphi(x-t-kT) \\ = \;& \sum_{n\in\mathbf{Z}} m(nT)\,\varphi(x-nT) \end{aligned}$$

by support considerations on the function $r(t)$. Therefore we have the summation formula

$$m(x) = \sum_n m(nT)\,\varphi(x-nT)$$

where the sum converges unconditionally in $M^{p,1}$ if $1 \le p < \infty$ and weak-* if $p = \infty$, and moreover there are constants $0 < A \le B$ such that for all such f,

$$A\|f\|_{M^{p,1}} \le \|\{f(nT)\}\|_{\ell^p} \le B\|f\|_{M^{p,1}}.$$

Taking $p = 2$, this recovers the classical sampling formula when the sampling is above the Nyquist rate.

4. Spreading functions with nonrectangular support and Bello's conjecture

In 1969, P. A. Bello [1] argued that what is important for channel identification is not the product ab of the maximum time-delay and Doppler shift of the channel but the area of the support of the spreading function. It is notable that Kailath also asserted something along these lines. This means that a time-variant channel whose spreading function has essentially arbitrary support is identifiable as long as the area of that support is smaller than one.
Using ideas from [6], Bello's conjecture was proved in [9].

Theorem 5. $\mathcal{H}_S$ is identifiable if $\mathrm{vol}^+(S) < 1$, and not identifiable if $\mathrm{vol}^-(S) > 1$. Here $\mathrm{vol}^+(S)$ is the outer Jordan content and $\mathrm{vol}^-(S)$ the inner Jordan content of S. In this case, the channel is identified by $g = \sum_n c_n\,\delta_{n/L}$ where $L \in \mathbf{N}$ and the L–periodic sequence $\{c_n\}$ is chosen based on the geometry of S.

We next present a generalization of Theorem 4 to this case. Before stating the result, a few preliminaries are required.

Definition 1. Given $L \in \mathbf{N}$, let $\omega = e^{-2\pi i/L}$ and define the *translation operator* T on $(x_0, \ldots, x_{L-1}) \in \mathbf{C}^L$ by

$$Tx = (x_{L-1}, x_0, x_1, \ldots, x_{L-2}),$$

and the *modulation operator* M on $\mathbf{C}^L$ by

$$Mx = (\omega^0 x_0, \omega^1 x_1, \ldots, \omega^{L-1} x_{L-1}).$$

Given a vector $c \in \mathbf{C}^L$ the *finite Gabor system with window* c is the collection $\{T^q M^p c\}_{q,p=0}^{L-1}$.

Note that the discrete Gabor system defined above consists of L^2 vectors in $\mathbf{C}^L$ so is necessarily overcomplete.

Definition/Proposition 2. The Zak Transform is defined for $f \in \mathcal{S}(\mathbf{R})$ by $Zf(t,\nu) = \sum_n f(t-n)\,e^{2\pi i n\nu}$. $Zf(t,\nu)$ satisfies the quasi-periodicity relations $Zf(t+1,\nu) = e^{2\pi i\nu}\,Zf(t,\nu)$ and $Zf(t,\nu+1) = Zf(t,\nu)$. Z can be extended to a unitary operator from $L^2(\mathbf{R})$ onto $L^2([0,1]^2)$.

If the spreading function of H, $\eta_H(t,\nu)$, is supported in a bounded Jordan region $S \subseteq \mathbf{R} \times \widehat{\mathbf{R}}$ with $\mathrm{vol}^+(S) < 1$, then by appropriately shifting and scaling η_H we can assume without loss of generality that for some $L \in \mathbf{N}$, $S \subseteq [0,1] \times [0,L]$ and that S meets at most L of the L^2 rectangles $R_{q,m} = ([0,1/L] \times [0,1]) + (q/L, m)$, $0 \le q,\, m < L$ whose union is $[0,1] \times [0,L]$. We can further assume that S does not meet any of the rectangles $R_{q,m}$ on the "edge" of the larger rectangle, specifically it does not meet $R_{q,m}$ with $q = 0$, $m = 0$, $q = L-1$ or $m = L-1$. The following Lemma connects the output $Hg(x)$ where $g = \sum_n c_n\,\delta_{n/L}$ to the spreading function $\eta_H(t,\nu)$. From this a reconstruction formula analogous to that in Theorem 4 can be derived.

Lemma 1. Given a period-L sequence (c_n) and $g = \sum_n c_n\,\delta_{n/L}$, then for (t,ν) in a sufficiently small neighborhood of $[0,1/L] \times [0,1]$,

$$\begin{aligned} & e^{-2\pi i\nu p/L}\,(Z \circ H)g(t+p/L,\nu) \\ = \;& \sum_{q=0}^{L-1}\sum_{m=0}^{L-1} (T^q\,M^m c)_p\,e^{-2\pi i\nu q/L}\,\eta_H(t+q/L,\nu+m). \end{aligned}$$

In other words, the spreading function can be realized as coefficients on the vectors of a finite Gabor system. The system is in general underdetermined since there are L

equations and L^2 unknowns. If, however, the support set S of the spreading function $\eta_H(t,\nu)$ satisfies $\mathrm{vol}^+(S) < 1$ and since S meets at most L of the rectangles $R_{q,m}$, there are at most L nonzero unknowns in the above linear system. If the resulting $L \times L$ matrix is invertible, then η_H can be determined uniquely from Hg. The vector c must be chosen so that this matrix is invertible. It is shown in [7] that if L is prime then such a c always exists.
We can prove the following theorem (cf. [8], [9]).

Theorem 6. Let $1 \leq p, q \leq \infty$. If $\mathrm{vol}^-(S) > 1$ then $OPW^{p,q}(S)$ is not identifiable. If $\mathrm{vol}^+(S) < 1$ then $OPW^{p,q}(S)$ is identifiable via operator sampling, and the identifier is of the form $g = \sum_n c_n \delta_{n/L}$ where $L \in \mathbf{N}$ and (c_n) is an appropriately chosen period-L sequence. Moreover, we have the formula

$$h_H(t,x) = \sum_{j=0}^{L-1} r_j(t) \sum_{k\in\mathbf{Z}} b_{j,k}\,(Hg)(t - q_j/L + k/L) \times \varphi_j(x - t - q_j/L - k/L)$$

unconditionally in $M^{1p,q1}$ and in the weak-* sense if $p = \infty$ or $q = \infty$. For $0 \leq j < L$, the rectangles R_{q_j,m_j} are precisely those that meet S. Also for each $0 \leq j < L$, $r_j(t)\widehat{\varphi}_j(\nu) = 1$ on R_{q_j,m_j} and vanishes outside a small neighborhood of R_{q_j,m_j}, and $b_{j,k}$ is a period-L sequence in k based on the inverse of the matrix derived from the discrete Gabor system that appears in Lemma 1.

5. Conclusion

This paper contains a brief overview of some recent results on the measurement and identification of communication channels and the relation of these results to sampling theory. These connections provide explicit reconstruction formulas for identification of operators modelling time-variant linear channels.

References:

[1] P.A. Bello. Measurement of random time-variant linear channels. 15:469–475, 1969.

[2] K. Gröchenig. *Foundations of Time-Frequency Analysis*. Applied and Numerical Harmonic Analysis. Birkhäuser, Boston, MA, 2001.

[3] T. Kailath. Sampling models for linear time-variant filters. Technical Report 352, Massachusetts Institute of Technology, Research Laboratory of Electronics, 1959.

[4] T. Kailath. Measurements on time–variant communication channels. 8(5):229–236, Sept. 1962.

[5] T. Kailath. Time–variant communication channels. *IEEE Trans. Inform. Theory: Inform. Theory. Progress Report 1960–1963*, pages 233–237, Oct. 1963.

[6] W. Kozek and G.E. Pfander. Identification of operators with bandlimited symbols. *SIAM J. Math. Anal.*, 37(3):867–888, 2006.

[7] J. Lawrence, G.E. Pfander, and D. Walnut. Linear independence of Gabor systems in finite dimensional vector spaces. *J. Fourier Anal. Appl.*, 11(6):715–726, 2005.

[8] G. Pfander and D. Walnut. On the sampling of functions and operators with an application to Multiple–Input Multiple–Output channel identification. In Manos Papadakis Dimitri Van De Ville, Vivek K. Goyal, editor, *Proc. SPIE Vol. 6701, Wavelets XII*, pages 67010T–1 – 67010T–14, 2007.

[9] G.E. Pfander and D. Walnut. Measurement of time–variant channels. *IEEE Trans. Inform. Theory*, 52(11):4808–4820.

[10] G.E. Pfander and D. Walnut. Operator identifcation and Feichtinger's algebra. *Sampl. Theory Signal Image Process.*, 5(2):151–168, 2006.

Special session on

Sampling
and
Industrial Applications

Chair: Laurent FESQUET

An Event-Based PID Controller With Low Computational Cost

Sylvain Durand and Nicolas Marchand

NeCS Project-Team, INRIA - GIPSA-lab - CNRS, Grenoble, France.
sylvain.durand@inrialpes.fr, nicolas.marchand@gipsa-lab.inpg.fr

Abstract:

In this paper, some improvements of event-based PID controllers are proposed. These controllers, contrary to a time-triggered one which calculates the control signal at each sampling time, calculate the new control signal only when the measurement signal *sufficiently* changes. The contribution of this paper is a low computational cost scheme thanks to a minimum sampling interval condition. Moreover, we propose to reduce much more the error margin during the steady state intervals by adding some extra samples just after transients. A cruise control mechanism is used for simulations and a noticeable reduction of the mean control computation cost is finally achieved with similar closed-loop performances to the conventional time-triggered ones.

1. Introduction

The classical so-called discrete time framework of controlled systems consists in sampling the system uniformly in the time with some constant sampling period h_{nom} and in computing and updating the control law every time instants $t = kh_{nom}$. This field, denoted time-triggered (or synchronous in sense that all the signal measurements are synchronous), has been widely investigated [6] even in the case of sampling jitter or measure loss that can be seen as some asynchronicity. However, some works addressed more recently event-based sampling where the sampling intervals are event-triggered (also called asynchronous), as for example when the output crosses a certain level. Thus the term *sampling period* denotes a time interval between two consecutive level crossings and the sampling periods are hence not equidistant in time anymore.

Event-triggered notion is taking more and more importance in the signal processing community with now various publications on this subject (see for instance [1] and the references therein). In the control community, very few works have been done. In [3], it is proved that such an approach reduces the number of sampling instants for the same final performance. In [8], it is shown that controlling an asynchronous sampled system or a continuous time system with quantized measurements and a constant control law over sampling periods are equivalent problems.

Many reasons are motivating event-based systems and in particular because more and more asynchronous systems or systems with asynchronous needs are encountered. Actually, the demand of low power electronic components in all embedded applications encourages companies to develop asynchronous versions of the existing time-triggered components, where a significant power consumption reduction can be achieved by decreasing the samplings and consequently the CPU utilization: about four times less power than its synchronous counterpart for the 80C51 microcontroller of Philips Semiconductors in [12]. Note that the sensors and the actuators based on level crossing events also exist, rendering a complete asynchronous control loop now possible. But the most important contributions come from the real-time control community. Indeed, real-time synchronous control tasks are often considered as hard tasks in term of time synchronization, requiring strong real time constraints. Efforts are so carried on the co-design between the controller and the task scheduler in order to soften these constraints. The adopted approach is often either to change dynamically the sampling period related to the load [10, 11] or to use event-driven control where the events are generated with a mix of level crossings and a maximal sampling period [9, 2].

This maximal sampling period seems to be added for stability reasons in order to fulfill the condition of Nyquist-Shannon sampling theorem: a new control signal is performed when the time elapsed since the last sample exceeds a certain limit. We first proposed in [7] to remove it because, thanks to the level detection, the Nyquist-Shannon sampling condition is no more consistent. The CPU cost is hence considerably reduced without performance loss. We now focus on the improvement of event-based control by reducing even more the computational cost with a controller based on a fully asynchronous level detection. The next two sections recall the conventional time-triggered structure and the existing event-based algorithms. The main contribution is developed in section 4 where an event-driven controller with low computational cost is detailed. All controllers are finally compared (in terms of performances and CPU needs) in section 5.

Notations:
e^- will denote the value of e at the last sampling time.

2. Time-Based Control

The textbook PID controller is given as follows:

$$U(s) = K\left(E(s) + \frac{1}{T_i s}E(s) + T_d s E(s)\right)$$

This equation can be divided into a proportional, an integral and a derivative parts, i.e. U_p, U_i and U_d respectively, which are then modified to improve performances [4]. First, set point weighting is applied on U_p and U_d for a more flexible structure, giving the PID two dimensions of freedom. Moreover, a low-pass filter is added in the derivative term to avoid problems with high frequency measurement noise.

$$\begin{aligned} U_p(s) &= K\left(\beta Y_{sp}(s) - Y(s)\right) \\ U_i(s) &= \frac{K}{T_i s} E(s) \\ U_d(s) &= \frac{K T_d s}{1 + T_d s/N}\left(\gamma Y_{sp}(s) - Y(s)\right) \end{aligned}$$

A discrete time controller is finally obtained: the proportional part is straightforward and the backward difference approximation is used for integral and derivative parts.

3. Event-Based Control

The basic setup of an event-based PID controller, introduced in [2], consists of two parts: *a time-triggered event detector* used for level crossings and *an event-triggered PID controller* which calculates the control signal. The first part runs with the sampling period h_{nom} (that is the same as for the corresponding conventional time-triggered PID) whereas the second part runs with the sampling interval h_{act} which depends on the requests sent by the event detector when a new control signal has to be calculated. This is required either when the relative error between the measured signal and the desired one crosses a certain level, i.e. $abs(e - e^-) > e_{lim}$, or if the maximal sampling period is achieved, i.e. $h_{act} \geq h_{max}$.

We proposed in [7] to remove this maximal sampling period underlying a primordial fact in asynchronous control that is that the Nyquist-Shannon sampling condition is no more consistent thanks to the level detection. However, the integral part, i.e. $u_i = u_i^- + K/T_i \cdot h_{act} \cdot e$, leads to important overshoots after the steady states since the sampling period h_{act} becomes huge due to the absence of event. In fact, this time interval between *the last sample before the steady state* and *the first sample of the transient* can be divided into a "real" steady state interval which is equal to $h_{act} - h_{nom}$, plus the detection time period h_{nom}. During the first part the error is very small (lower than e_{lim} else the steady state is not achieved) and so is the product he (lower than $(h_{act} - h_{nom})\, e_{lim}$). As regards the second part, when the set point changes the error becomes large but only during the event detection and therefore the product is $h_{nom}e$. From this observation, several control algorithms were proposed in [7] and we will use the hybrid one which gives good performances with the minimum of samplings.
The hybrid algorithm is a mix between **i)** a controller with a saturation of he which is bounded in $(h_{act} - h_{nom}) \cdot e_{lim} + h_{nom} \cdot e$ when $h_{act} \geq h_{max}$ and **ii)** a controller with an exponential forgetting factor of h_{act} to decrease its impact after a long steady state interval, with $h_{act}^i = h_{act} \cdot \exp\left(h_{nom} - h_{act}\right)$ corresponding to the new sampling period used in the integral part. This mix leads to bound the exponential forgetting factor:

$$\begin{aligned} &\text{if } h_{act} \geq h_{max} \\ &\quad he = \left(h_{act}^i - h_{nom}\right) \cdot e_{lim} + h_{nom} \cdot e \\ &\text{else} \\ &\quad he = h_{act} \cdot e \\ &\text{end} \\ &u_i = u_i^- + K/T_i \cdot he \end{aligned} \tag{1}$$

A first improvement could be obtained by changing the level crossing detection since only one level is really required. Indeed, the control signal needs to be calculated when the measurement is too far from the set point, i.e. as soon as $abs(e) > e_{lim}$. Of course, with this method the number of samples increases during the transients but, at least, the error between the system and the set point is now sure to be lower than e_{lim} during the steady state intervals, which was not the case before with the level detection of the relative error $abs(e - e^-) > e_{lim}$.
A second improvement could be done on the time-triggered event detector which is currently a discrete time system: an event could only be detected at the time instants $t = kh_{nom}$ thereof several levels could miss if they appear between two sampling instants. Thus we propose to use a continuous time event detector which is in fact closer to the real case since a sensor based on level crossing events will send a request as soon as a level is crossed.

Afterwards, the hybrid controller with these improvements is called the asynchronous event-based controller.

4. Event-Based Control with Low Computational Cost

The asynchronous event-based controller is interesting but the number of samples is still important during transients. Indeed, a new request is sent as soon as the error is upper than the detection limit, i.e. $abs(e) > e_{lim}$, which means (quasi)-continuously during the whole transient. To avoid that, we propose to add a minimum sampling interval condition to lighten the transients in order that a new control signal is performed only if a certain time was elapsed since the last sample, i.e. $h_{act} > h_{min}$. This minimum sampling interval could be chosen as the discrete sampling period h_{nom} corresponding to the conventional time-triggered controller or not, but it does have to satisfy the Nyquist-Shannon sampling condition. The choice $h_{min} = h_{nom}$ leads to a discrete-time event detector when the dynamics is important and to a continuous-time event detector when the dynamics is slow (quasi-steady state). Thus, when an event occurs after a steady state configuration, a new control signal is instantaneously computed.

Whatever that may be the h_{min} value, an important reduction of the computational cost is achieved. Nevertheless, we propose to improve the event-based scheme again by adding a few number of samples more. The idea here is to decrease much more the error during the steady state intervals. Currently, one could assure that the error is lower than the limit e_{lim} but cannot know how much lower. Moreover, one could not know if the measured signal is going closer or moving away from the set point.

Therefore, we propose to add some extra samples after a transient while an event-based controller would do not do anything because the condition $abs(e) > e_{lim}$ is wrong. Thus, an extra event is sent to the controller if nothing appends after the last time a control signal was calculated plus a certain sampling interval h_{extra}. Then, this is repeated while the error is upper than a desired minimum level e_{min}. One only needs to define his desired error margin and some extra samples will be added to achieve that. Note that the lower e_{min} is chosen the higher the number of extra samples will be.

5. Simulation Results: Application to a Cruise Control Mechanism

Event-based controller is a good solution, more especially for all the systems which do not need to be constantly controlled. We chose to illustrate our proposals with the cruise control mechanism depicted in [5] because the desired speed of the car is constant most of the time and a new control signal is so only required when the set point changes or when the load (i.e. the slope of the road) varies.

The equation of motion of the car (ν is the velocity) is:

$$m\dot{\nu} = F - F_d$$

The force F is generated by the engine, whose torque is proportional to a control signal $0 \leq u \leq 1$ that controls the throttle position and depends on engine velocity too.

$$F = \alpha_n u T_m \left(1 - \beta \left(\frac{\alpha_n \nu}{\omega_m} - 1\right)^2\right)$$

where α_n depends on the gear ratio n.
The disturbance force F_d has three major components due to the gravity F_g, to the rolling friction F_r and to the aerodynamic drag F_a.

$$\begin{aligned} F_d &= F_g + F_r + F_a \\ \text{with} \quad F_g &= mgsin(\theta) \\ F_r &= mgC_r sgn(\nu) \\ F_a &= \tfrac{1}{2}\rho C_d A \nu^2 \end{aligned}$$

where θ is the slope of the road, i.e. the disturbance.

As regards the control law, an anti-windup mechanism is added to consider the saturation of the control signal u. Thus the integral part consists on the integral of the error plus a *reset* based on the saturation of the actuator (in order to prevent windup when the actuator is saturated).

$$u_i = u_i^- + \frac{K}{T_i}x - \frac{h_{act}}{T_a}(u - u_{sat})$$

where $x = h_{act} \cdot e$ for the time-triggered controller and $x = he$ defined by (1) for the event-based controllers.
Parameter values are $K = 0.8$, $T_i = 1.4$ and $T_a = 0.7$. The nominal and maximal sampling intervals used for the hybrid algorithm are $h_{nom} = 0.1s$ and $h_{max} = 0.5s$ and those used for the low computational cost and the extra samples ones are $h_{min} = 0.1s$ and $h_{extra} = 0.5s$. The detection levels are $e_{lim} = 0.1$ and $e_{min} = 0.01$ for crossing events and for extra samples respectively.

The simulations run during $50s$ with the following test bench: at time 0 the set point is set to $25m/s$ $(90km/h)$, then at time $2s$ it is changed to $30.6m/s$ $(110km/h)$ and changed again to $36.1m/s$ $(130km/h)$ at time $30s$. The gear ratio is chosen accordingly to the speed range, i.e. $n = 5$, and no disturbance is applied, i.e. $\theta = 0$.

The first simulation results are shown on Figure 1 where the conventional time-triggered PI controller is compared to the asynchronous event-based one (see section 3). The top plot shows the set point and both measured signals, the bottom plot shows the sampling intervals (i.e. this signal changes each time the controller calculates a new control signal). The asynchronous event-based controller permits to obtain a system response as quick as the time-triggered one, by calculating a control signal about four time less only (with this benchmark). However, the number of samples remains important during the transients. Our proposal, i.e. the event-based PI controller with a low computational cost, avoids that since the number of samples is dropped by a ratio of 30, as shown on Figure 2.

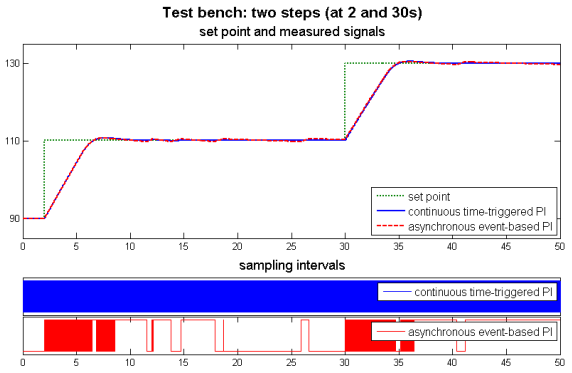


Figure 1: *A conventional time-triggered PI controller* (15000 sampling intervals) vs. *the asynchronous event-based one* (3703 sampling intervals, that is 24.7%).

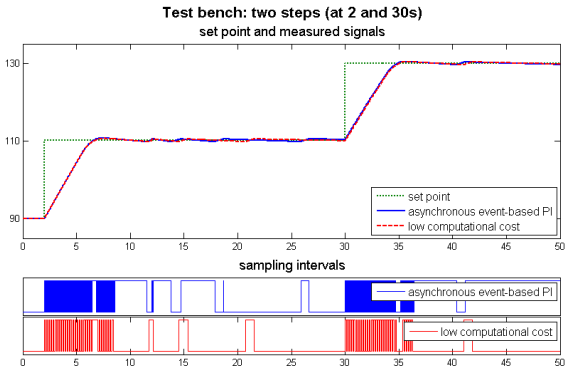


Figure 2: *The asynchronous event-based PI controller* (3703 sampling intervals) vs. *the one with a low computational cost* (126 sampling intervals, that is 3.4%).

Whatever the achieved gain with the low computational cost controller, we propose to improve the error during the steady state intervals by adding some samples just after the transients. Results are shown on Figure 3 where one could see that, by adding extra samples, the sampling number is

finally reduced and the steady state intervals are not oscillating anymore. These are thanks to a measurement signal closer to the set point during the steady state intervals.

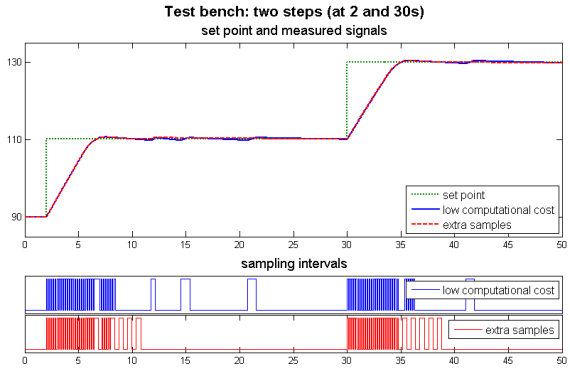


Figure 3: *The asynchronous event-based PI controller with a low computational cost* (126 sampling intervals) vs. *the one with extra samples* (120 sampling intervals).

Finally the integral of the norm of the error are compared for the whole controllers to verify if the responses are not too far from the conventional time-triggered one. All measurements on Figure 4 have a similar behavior with some differences during the steady state intervals because of the allowed error margin e_{lim}. The final values are 74.67 for the reference, 78.2 for the asynchronous event-based controller, 78.63 for the low computational cost one and 77.12 for the extra samples one. Moreover, as regards the last one, it is possible to be much more closer to the time-based value by reducing the minimum value e_{min}.

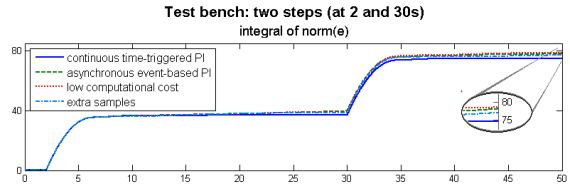


Figure 4: Integral of the norm of the error.

6. Conclusions and Future Works

In this paper we propose to improve the event-based PID controllers depicted in [2] and [7]. The first improvement consists on a minimum sampling interval condition used to decrease the number of samples during the transients. The second one comes from the wishing to reduce much more the error margin during the steady state intervals. Based on these ideas, event-based PID controllers with low computational cost and with extra samples are proposed.
A cruise control mechanism is used to compare them (in simulation) with the conventional time-triggered and with the classical event-based controllers. Both proposals clearly give good performances with a minimum of sampling intervals and the controller with extra samples permits to reduce the error margin as low as desired to achieve a response very closed to the conventional one.

Next steps in this research is naturally to test these controllers in practice and develop other event-based methods for more general types of control.

7. Acknowledgments

This research has been supported by the GIPSA-lab, CNRS and INRIA in the FeedNetBack project context. The project aims to close the control loop over wireless networks by applying a co-design framework that allows the integration of communication, control, computation and energy management aspects in a holistic way.

References:

[1] F. Aeschlimann, E. Allier, L. Fesquet, and M. Renaudin. Asynchronous FIR filters: towards a new digital processing chain. In *Proceedings of the 10th International Symposium on Asynchronous Circuits and Systems*, pages 198–206, 2004.

[2] K-E Årzén. A simple event-based PID controller. In *Preprints of the 14th World Congress of IFAC*, 1999.

[3] K.J. Åström and B. Bernhardsson. Comparison of Riemann and Lebesque sampling for first order stochastic systems. In *Proceedings of the 41st IEEE Conference on Decision and Control*, 2002.

[4] K.J. Åström and T. Hägglund. *PID controllers: theory, design, and tuning, 2nd Edition*. The Instrumentation, Systems, and Automation Society, 1995.

[5] K.J. Åström and R.M. Murray. *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton University Press, 2008.

[6] K.J. Åström and B. Wittenmark. *Computer Controlled Systems, 3rd Edition*. Prentice Hall, 1997.

[7] S. Durand and N. Marchand. Further results on event-based PID controller. In *Proceedings of the European Control Conference*, 2009.

[8] N. Marchand. Stabilization of Lebesgue sampled systems with bounded controls: the chain of integrators case. In *Proceedings of the 17th IFAC World Congress*, 2008.

[9] J.H. Sandee, W. Heemels, and P.P.J. van den Bosch. Event-driven control as an opportunity in the multidisciplinary development of embedded controllers. In *Proceedings of American Control Conference*, pages 1776–1781, 2005.

[10] O. Sename, D. Simon, and D. Robert. Feedback scheduling for real-time control of systems with communication delays. In *Proceedings of the IEEE Conference on Emerging Technologies and Factory Automation*, volume 2, 2003.

[11] D. Simon, D. Robert, and O. Sename. Robust control/scheduling co-design: application to robot control. In *Proceedings of the IEEE Symposium on Real-Time and Embedded Technology and Applications*, pages 118–127, 2005.

[12] H. van Gageldonk, K. van Berkel, A. Peeters, D. Baumann, D. Gloor, and G. Stegmann. An asynchronous low-power 80C51 microcontroller. In *Proceedings of the 4th International Sympsonium on Advanced Research in Asynchronous Circuits and Systems*, pages 96–107, 1998.

A coherent sampling-based method for estimating the jitter used as entropy source for True Random Number Generators

Boyan Valtchanov, Viktor Fischer, Alain Aubert

Laboratoire Hubert Curien UMR CNRS 5516, Bât. F 18 Rue du Professeur Benoît Lauras , 42000 Saint Etienne, France.
{boyan.valtchanov,fischer,alain.aubert}@univ-st-etienne.fr

This paper was partially supported by the Rhône-Alpes Region and Saint-Etienne Métropole, France

Abstract:

The paper presents a method, which can be employed to measure the timing jitter present in periodic clock signals that are used as entropy source in true random number generators aimed at cryptographic applications in reconfigurable hardware. The method uses the principle of a coherent sampling and can be easily implemented inside the chip in order to test online the jitter source. The method was carefully validated in various simulations that have shown that the measured jitter size corresponds perfectly to that of the jitter injected to the model. While the primary aim of the proposed measuring technique was the evaluation of the quality of jitter as an entropy source in random number generators, we believe that the same principle can be used in order to characterize the jitter in fast communication links as well.

1. Introduction

In the global communication era, more and more recent industrial applications need to secure data and communications. Many cryptographic primitives and protocols that are used to ensure confidentiality, integrity and authenticity use random number generators in order to generate confidential keys, initial vectors, nonces, padding values, etc. While random bit-stream generators can be easily implemented in analog or mixed-signal devices, the generation of random bit-streams is a challenging task when the generator should be implemented in a logic device like FPGAs (Field-Programmable Gate Arrays). Clearly, logic devices are well suited for algorithmic (pseudo) random number generators, but the true-random number generators need sources of randomness that are difficult to find and explore in logic devices. A mathematical model of the true random number generator (TRNG) is also a crucial element of the cryptographic application design since the final entropy of the generated random bit-stream could be characterized and thus certified if one is able to characterize the physical phenomenon that is used as the entropy source. If the model does not exist, there would be no guarantee that the final entropy of the output stream is true-random, pseudo-random or perhaps a mixture of random and pseudo random phenomena. Characterizing and monitor the entropy source (the jitter) and proposing a mathematical model is the main motivation of the paper.

2. Jitter as an entropy source for TRNGs

Many of the TRNGs known up to date [1], [4], [5], use the jitter present in clock signals (generated using ring oscillators, phase-locked loops or delay-locked loops) as a source of entropy. The quality of the generated random bits is related to the parameters of the clock jitter. In order to avoid jitter manipulations and attacks, it is important to measure these parameters on-line and, if possible, inside the device.

The jitter can be defined as a short-term variation of an event from its ideal position [6]. In general, it is expressed as the variation in time of the zero crossing (rising or falling edge) of the clock signal. The jitter can be a good candidate for randomness generation, since its behavior is closely related to the thermal noise inside semiconductors [2]. The advantage of the thermal noise employed as a source of randomness is that it is relatively difficult to manipulate it in order to realize an attack on the TRNG. The method presented in this paper considers only a true-random (Gaussian) jitter component and it does not take into account the deterministic behavior of the jitter at this stage of our research. For a deeper understanding of the jitter behavior we recommend to read [9].

3. Principle and theoretical background

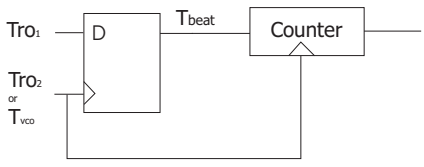


Figure 1: Random jitter component measurement based on the coherent sampling.

The proposed method allows to accurately quantify the random component of the jitter present in clock signals generated inside logic devices. Although the technique can be used to measure the jitter, it has been developed not for measurement or testing purposes, but rather for modeling a TRNG that uses the jitter as a source of randomness.

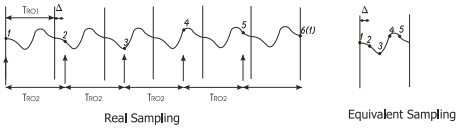


Figure 2: Principle of the coherent sampling.

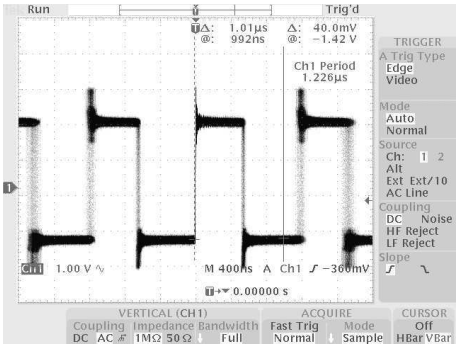


Figure 3: Experimental T_{Beat} signal example.

The proposed measurement technique (see Figure 1) is based on a coherent sampling: the sampling of a periodic signal by another periodic signal featuring similar frequency [3]. The signal on the output of the sampler is called a beat signal and it is a low-frequency signal depending on the frequency difference Δ between the two clock signals T_{ro_1} and T_{ro_2}.

Figure 2 shows the principle of the coherent sampling using two (clock) signals having similar frequencies and the resulting beat signal T_{Beat}, representing the image of T_{ro_1}. An example of this T_{Beat} signal captured on oscilloscope is given in Figure 3. Using the infinite persistence of the oscilloscope, we can clearly see the variations of the period of the beat signal. These variations are the consequence of the jitter present in T_{ro_1} and T_{ro_2} signals. Because of the coherent relationship between the two frequencies, each "half-period" of the beat signal is an integer number of the clock period T_{ro_2}. A counter clocked with this clock signal can thus be used in order to represent these variations. In the next section, we will discuss how we can compute the jitter present in T_{ro_1} by observing the variations in a population of several T_{Beat} periods.

If the proposed technique would be used to measure precisely the jitter of the internal clock signal, one should use an accurate external low phase-noise VCO (Voltage Controlled Oscillator) as a sampling clock and accurately tune its period in relationship to the internal clock period in order to obtain a small Δ. Instead, in order to model the TRNG behavior and to measure the jitter inside the device, we have used two ring oscillators, implemented in the same FPGA. Both oscillators have the same number of inverters. In order to guarantee a small difference between clock periods (Δ), the placement and routing have to be done manually. The final period difference is thus caused mainly by the different delays of the routing scheme selected by the placement and routing tool. Next, we will analyze the case, when only random (Gaussian) jitter component is present in the generated clock signals.

3.1 Measurement of the true-random jitter component

Let us assume that the two clock signals are derived from two internal ring oscillators, and let $T_{ro_{1_{Ideal}}}$ and $T_{ro_{2_{Ideal}}}$ be the two ideal jitter-free periods. We need to achieve a small time period difference between $T_{ro_{1_{Ideal}}}$ and $T_{ro_{2_{Ideal}}}$, namely:

$$T_{ro_{2_{Ideal}}} = T_{ro_{1_{Ideal}}} + \Delta_{Ideal}. \quad (1)$$

This difference comes from the fact that even with the same number of delay elements the two ring oscillators differs due to process variations during manufacturing. With a careful placement, one can obtain Δ of several tens of picoseconds. However the Δ wont be reproducible from one chip to another.
If a random jitter would be included in the previous equations, we obtain:

$$T_{ro_1} = T_{ro_{1_{Ideal}}} + N(0, \sigma_1) = N(T_{ro_{1_{Ideal}}}, \sigma_1) \quad (2)$$

$$T_{ro_2} = T_{ro_{2_{Ideal}}} + N(0, \sigma_2) = N(T_{ro_{2_{Ideal}}}, \sigma_2) \quad (3)$$

Where $N(0, \sigma)$ denote a zero-mean Normal distribution with standard deviation σ.
We can then express the difference Δ by:

$$\Delta = N(T_{ro_{2_{Ideal}}}, \sigma_2) - N(T_{ro_{1_{Ideal}}}, \sigma_1) \quad (4)$$

$$\Delta = N(\Delta_{Ideal}, \sqrt{\sigma_1^2 + \sigma_2^2}) \quad (5)$$

If σ_1 is the same as σ_2, what is the case when the two signals are derived from internal ring oscillators, we get

$$\Delta = N(\Delta_{Ideal}, \sqrt{2}\sigma) \quad (6)$$

Otherwise one should make precise characterization of the VCO used to match the frequencies in order to measure the σ_{VCO}.
According to [8], we can express the length of T_{Beat} as:

$$\frac{T_{Beat}}{\Delta_{Ideal}} = N(\frac{T_{ro_{1_{Ideal}}}}{\Delta_{Ideal}}, \sqrt{\frac{T_{ro_{1_{Ideal}}}}{\Delta_{Ideal}}}\sqrt{\sigma_1^2 + \sigma_2^2}) \quad (7)$$

which, if σ_1 equals σ_2, simplifies to:

$$\frac{T_{Beat}}{\Delta_{Ideal}} = N(\frac{T_{ro_{1_{Ideal}}}}{\Delta}, \sqrt{\frac{T_{ro_{1_{Ideal}}}}{\Delta_{Ideal}}}\sqrt{2}\sigma) \quad (8)$$

The length of the resulting beat signal, T_{Beat} can be then expressed as a normal process:

$$\frac{T_{Beat}}{\Delta_{Ideal}} = N(\mu_{T_{Beat}}, \sigma_{T_{Beat}}) \quad (9)$$

with the mean and standard deviation:

$$\mu_{T_{Beat}} = \frac{T_{ro_{1_{Ideal}}}}{\Delta}, \sigma_{T_{Beat}} = \sqrt{\frac{T_{Ro_{Ideal}}}{\Delta_{Ideal}}}\sqrt{2}\sigma \quad (10)$$

In consequence, if we measure the μ_{T_Beat} and $\sigma_{T_{Beat}}$ using the principle presented in Figure 1, which is based on

the use of an 8-bit counter, we can precisely calculate the amount of the random jitter, expressed in 1σ ps, i.e. the RMS jitter (root mean square) present in the two clock signals using equation (11).

$$\sigma = \frac{\sigma_{T_{Beat}} \Delta_{Ideal}}{\sqrt{\frac{T_{Ro_{Ideal}}}{\Delta_{Ideal}}\sqrt{2}}} \quad (11)$$

4. Simulation results

In order to validate equation (11), we have used a simulation model presented in [8] and depicted in Figure 4. The random jitter is generated in text files using Matlab and then injected in VHDL simulation using the Textio package. We have injected different amounts of random jitter (RMS) to the clock signals and analyzed the obtained values of the counter. The $T_{ro_{1_{Ideal}}}$ was set to 5 ns (200Mhz) and Δ to 40 ps. The results of the simulations and recalculated jitter values using equation 11 are presented in Table 4. As it can be seen, the measurement precision that can be achieved is close to 1 ps RMS. Figure 5 present the case for 7 ps RMS jitter present in both T_{ro_1} and T_{ro_2} signals.

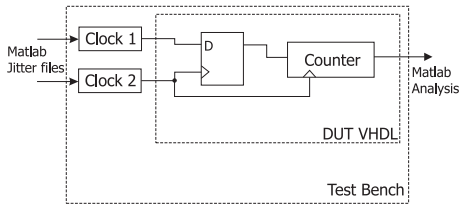


Figure 4: Simulation setup.

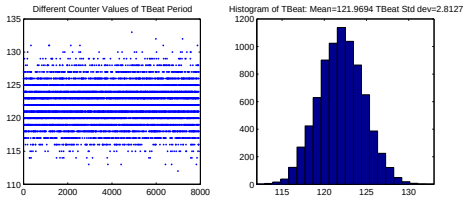


Figure 5: Histogram of the simulated T_{Beat}.

Injected 1σ RMS jitter [ps]	Measured μ_{Tbeat}	Measured σ_{Tbeat}	Calculated 1σ RMS [ps]
10	121.93	4.03	10.19
9	121.98	3.64	9.20
8	121.97	3.24	8.19
7	121.97	2.81	7.10
6	121.98	2.47	6.24

Table 1: Simulation results of the random jitter quantification.

5. Discussion and conclusions

We have proposed a jitter measurement technique that can be embedded in FPGA devices for evaluating and monitoring of the source of randomness employed in true random number generators. The measurement technique can be used as well to characterize the jitter present in high-speed clock signals, if an external VCO (Voltage Controlled Oscillator) is used. The use of an external and precise clock source is necessary in order to closely match the period of the signal under test to the period of the reference clock signal. We have shown by simulation that the measurement error of the proposed method is less than 1 ps RMS of the random component of the jitter.
However, in real world situations and especially inside FPGAs, the jitter can exhibit a non negligible deterministic component due to various factors (power supply variations, cross-talks, R-F interference, etc...). In this case, equation (11) cannot be used for random component jitter quantification and the deterministic jitter has to be considered, too. However, we believe that it is possible to integrate this deterministic behavior of the jitter in the proposed model. This integration is the objective of our current research.

References:

[1] V. Fischer, M. Drutarovsky, M. Simka, and N. Bochard. High performance True Random Number Generator in Altera Stratix FPLDs. *Lecture notes in computer science, FPL'04*, pages 555–564, 2004.

[2] A. Hajimiri and TH Lee. A general theory of phase noise in electrical oscillators. *Solid-State Circuits, IEEE Journal of*, 33(2):179–194, 1998.

[3] J.L. Huang and K.T. Cheng. An On-Chip Short-Time Interval Measurement Technique for Testing High-Speed Communication Links. *Proceedings of the 19th IEEE VLSI Test Symposium*, page 380, 2001.

[4] P. Kohlbrenner and K. Gaj. An embedded true random number generator for FPGAs. *Proceedings of the 2004 ACM/SIGDA 12th international symposium on Field programmable gate arrays*, pages 71–78.

[5] B. Sunar, W.J. Martin, and D.R. Stinson. A Provably Secure True Random Number Generator with Built-In Tolerance to Active Attacks. *IEEE TRANSACTIONS ON COMPUTERS*, pages 109–119, 2007.

[6] T. Technologies. Synchronous Optical Network (SONET) Transport Systems: Common Generic Criteria. Technical report, GR-253-CORE, 2000.

[7] K.H. Tsoi, K.H. Leung, and P.H.W. Leong. Compact FPGA-based true and pseudo random number generators. *Field-Programmable Custom Computing Machines, 2003. FCCM 2003. 11th Annual IEEE Symposium on*, pages 51–61, 2003.

[8] B. Valtchanov, A. Aubert, F. Bernard, and V. Fischer. Modeling and observing the jitter in ring oscillators implemented in FPGAs. In *Design and Diagnostics of Electronic Circuits and Systems, 2008. DDECS 2008. 11th IEEE Workshop on*, pages 1–6, 2008.

[9] SW Wedge. Predicting random jitter-Exploring the current simulation techniques for predicting the noise in oscillator, clock, and timing circuits. *Circuits and Devices Magazine, IEEE*, 22(6):31–38, 2006.

Orthogonal Exponential Spline Pulses with Application to Impulse Radio

Masaru Kamada(1), Semih Özlem(2) and Hiromasa Habuchi(1)

(1) Ibaraki University, Hitachi, Ibaraki 316 8511, Japan.
(2) Bogazici University, Bebek, Istanbul, Turkey.
kamada@mx.ibaraki.ac.jp, semozl@gmail.com, habuchi@mx.ibaraki.ac.jp

Abstract:

With application to the impulse radio communications in mind, a locally supported and zero-mean pulse which is orthogonal to its shifts by integers is sought among the exponential splines having the knot interval $\frac{1}{2}$. An example pulse is obtained that complies with the regulation imposed by the US Federal Communications Commission and will potentially enable an impulse radio communications system as fast as 6G pulses per second.

1. Introduction

The M-shaped linear spline

$$M(t) = \begin{cases} \sqrt{3}t, & 0 \le t \le \frac{1}{2} \\ \sqrt{3}(2-3t), & \frac{1}{2} \le t \le 1 \\ \sqrt{3}(3t-4), & 1 \le t \le \frac{3}{2} \\ \sqrt{3}(2-t), & \frac{3}{2} \le t \le 2 \\ 0, & \text{elsewhere} \end{cases} \tag{1}$$

plotted in Fig. 1 is not a wavelet in the sense of muntiresolutional analysis because $M(t)$ is not orthogonal to its contracted version $M(2t)$. But it has three remarkable properties that (i) it is locally supported, (ii) its integration over the domain is zero, and (iii) its shifts by integers are orthogonal to one another [2]. Those properties are exactly what is required of pulses for the impulse radio communications [6]. The three properties are required (i) for the sake of real-time communications, (ii) for the pulse to be feasible as a radio waveform, and (iii) for pulse detection to be robust against noise in the sense of least-square estimation, respectively.

We shall look for this kind of pulse functions in the broader family of exponential splines [4, 5] which have the advantage that they can be shaped through linear dynamical systems [5] . The pulse functions, if they are found, will work as practical pulses which carry information in the impulse radio communications.

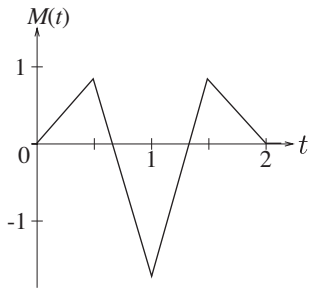


Figure 1: M-shaped linear spline.

The problem is simple: we are to find a locally supported and zero-mean exponential spline $q(t)$ with the knot interval $\frac{1}{2}$ that satisfies

$$\int_{-\infty}^{\infty} q(t)q(t-k)dt = \begin{cases} 1, & k=0 \\ 0, & k \neq 0 \end{cases} \tag{2}$$

for any ingeter k. This paper presents a procedure to find such a pulse function and its application to the impulse radio.

2. Construction of orthogonal pulses

Any exponential spline can be represented by a linear combination of the exponential B-spline and its shifts [4, 5]. An exponential B-spline with the knot interval $\frac{1}{2}$ is the output

$$\beta(t) = S(b)(t) \tag{3}$$

of a linear dynamical system S having the transfer function

$$G(s) = \frac{\mu_{n-1}s^{n-1} + \cdots + \mu_1 s + \mu_0}{(s-\lambda_0)(s-\lambda_1)\cdots(s-\lambda_{n-1})} \tag{4}$$

for the input being a series of delta functions

$$b(t) = \sum_{l=0}^{n} b_l \delta(t - l/2) \tag{5}$$

such that

$$\begin{aligned} B(z) &= \sum_{l=0}^{n} b_l z^{-\frac{l}{2}} \\ &= (1-z^{-\frac{1}{2}}\mathrm{e}^{\frac{\lambda_0}{2}})(1-z^{-\frac{1}{2}}\mathrm{e}^{\frac{\lambda_1}{2}})\cdots(1-z^{-\frac{1}{2}}\mathrm{e}^{\frac{\lambda_{n-1}}{2}}). \end{aligned} \tag{6}$$

This exponential B-spline is locally supported as

$$\beta(t) = 0, \ \ t \notin \left(0, \frac{n}{2}\right). \tag{7}$$

In order to keep the splines zero-mean, instead of the original exponential B-spline $\beta(t)$, we shall use

$$\alpha(t) = \beta(t) - \beta\left(t - \frac{1}{2}\right) \tag{8}$$

which has the zero mean

$$\int_{-\infty}^{\infty}\alpha(t)dt=0 \tag{9}$$

and is locally supported as

$$\alpha(t)=0,\ \ t\notin\left(0,\frac{n+1}{2}\right). \tag{10}$$

Another representation of this $\alpha(t)$ is the output

$$\alpha(t)=S(a)(t) \tag{11}$$

of S for the input

$$a(t)=\sum_{l=0}^{n}a_l\delta(t-l/2), \tag{12}$$

where

$$\begin{aligned}A(z)&=\sum_{l=0}^{n+1}a_lz^{-\frac{l}{2}}\\&=(1-z^{-\frac{1}{2}}\mathrm{e}^{\frac{\lambda_0}{2}})\cdots(1-z^{-\frac{1}{2}}\mathrm{e}^{\frac{\lambda_{n-1}}{2}})(1-z^{-\frac{1}{2}}).\end{aligned} \tag{13}$$

Let the desired pulse function be represented in the form

$$q(t)=\sum_{l=0}^{n-1}c_l\alpha(t-l/2). \tag{14}$$

Then it is automatic that $q(t)$ is locally supported as

$$q(t)=0,\ \ t\notin(0,n) \tag{15}$$

and has the zero mean

$$\int_{-\infty}^{\infty}q(t)dt=0. \tag{16}$$

The remaining request is that its autocorrelation

$$r(x)=\int_{-\infty}^{\infty}q(t)q(t-x)dt \tag{17}$$

should satisfy the orthogonality conditions

$$r(k)=\begin{cases}1, & k=0\\ 0, & k=\pm1,\pm2,\cdots\end{cases} \tag{18}$$

with respect to shift by integers. Here the number n of $\{\alpha(t-l/2)\}_{l=0}^{n-1}$ employed for composing $q(t)$ in (14) is chosen so that the number n of the unknown coefficients $\{c_l\}_{l=0}^{n-1}$ be the same as that of the essential conditions

$$r(k)=\begin{cases}1, & k=0\\ 0, & k=1,2,\cdots,n-1\end{cases} \tag{19}$$

reduced from (18) by (15) and the equality $r(x)=r(-x)$.

Now we have only to find the coefficients $\{c_l\}_{l=0}^{n-1}$ that make (19) hold good. Define

$$c(t)=\sum_{l=0}^{n-1}c_l\delta(t-l/2)\ \ \text{and}\ \ C(z)=\sum_{l=0}^{n-1}c_lz^{-\frac{l}{2}} \tag{20}$$

by $\{c_l\}_{l=0}^{n-1}$, and prepare time-reversed functions

$$\tilde{a}(t)=a(-t),\ \tilde{c}(t)=c(-t),\ \tilde{q}(t)=q(-t) \tag{21}$$

and the "mirror" system $\tilde{S}$ having the transfer function $G(-s)$. Then we can express the correlation by

$$\begin{aligned}r(k)&=(q*\tilde{q})(k)\\&=(S\circ\tilde{S})(a*\tilde{a}*c*\tilde{c})(k),\end{aligned} \tag{22}$$

where $*$ denotes the convolution integral, and we can write $D(z)=C(z)C(z^{-1})$ in the form

$$C(z)C(z^{-1})=d_0+\sum_{j=1}^{n-1}d_j(z^{-\frac{j}{2}}+z^{\frac{j}{2}}) \tag{23}$$

which implies

$$(c*\tilde{c})(t)=d_0\delta(t)+\sum_{j=1}^{n-1}d_j\left(\delta(t-j/2)+\delta(t+j/2)\right). \tag{24}$$

In the meantime, a locally supported exponential spline

$$\varphi(x)=(S\circ\tilde{S})(a*\tilde{a})(x) \tag{25}$$

associated with the composite system $S\circ\tilde{S}$ satisfies

$$\varphi(x)=\varphi(-x). \tag{26}$$

By (22), (24), (25) and (26), we can reduce the orthogonality conditions (19) to the linear equations

$$\begin{aligned}&d_0\varphi(k)+\sum_{j=1}^{n-1}d_j\left(\varphi(k-j/2)+\varphi(k+j/2)\right)\\&=\begin{cases}1, & k=0\\ 0, & k=1,2,\cdots,n-1.\end{cases}\end{aligned} \tag{27}$$

Solvability of (27) for $\{d_j\}_{j=0}^{n-1}$ can be checked by numerical computation in practice. A simpler condition in terms of dynamical parameters is yet to be established.

We assume that (27) is solvable since we cannot proceed unless this is the case. Then, $C(z)C(z^{-1})$ determined by (23) from $\{d_j\}_{j=0}^{n-1}$ can be factorized in the form

$$\begin{aligned}C(z)C(z^{-1})=\gamma_0&(z^{-\frac{1}{2}}-\gamma_1)(z^{\frac{1}{2}}-\gamma_1)(z^{-\frac{1}{2}}-\gamma_2)(z^{\frac{1}{2}}-\gamma_2)\\&\cdots(z^{-\frac{1}{2}}-\gamma_{n-1})(z^{\frac{1}{2}}-\gamma_{n-1}).\end{aligned} \tag{28}$$

Taking half the factors, we can find

$$C(z)=\pm\sqrt{\gamma_0}(z^{-\frac{1}{2}}-\gamma_1)(z^{-\frac{1}{2}}-\gamma_2)\cdots(z^{-\frac{1}{2}}-\gamma_{n-1}) \tag{29}$$

that gives the sought coefficients $\{c_l\}_{l=0}^{n-1}$ by (20). Exciting the system S with the input series of delta functions

$$v(t)=\sum_{l=0}^{n-1}c_la(t-l/2), \tag{30}$$

we obtain the desired pulse function

$$q(t)=S(v)(t)=\sum_{l=0}^{n-1}c_l\alpha(t-l/2). \tag{31}$$

In the case $G(s) = \frac{1}{s}$, the problem is trivial and the resulting pulse is the Haar function

$$H(t) = \begin{cases} 1, & 0 < t \le \frac{1}{2} \\ -1, & \frac{1}{2} < t \le < 1 \\ 0, & \text{elsewhere.} \end{cases} \quad (32)$$

The case $G(s) = \frac{1}{s^2}$ yields $M(t)$ of (1) as expected. Because it happens that $M(t) = \sqrt{3}(H * H)(t)$, we might speculate that the pulse associated with $G(s) = \frac{1}{s^3}$ could be proportional to $(H*H*H)(t)$. But that is not true since $(H * H * H)(t)$ is not orthogonal to $(H * H * H)(t-2)$. It is interesting as well as disappointing that we obtain a complex-valued pulse in the case $G(s) = \frac{1}{s^3}$. A nice example pulse will appear in the next section in the context of its application to the impulse radio communications.

3. Application to Impulse Radio

While the series of delta functions $a(t)$ does not exist in the real world, its integration

$$\int_{-\infty}^{t} a(\tau)d\tau = \begin{cases} 0, \ t < 0 \\ \sum_{k=0}^{l} a_k, \ \frac{l}{2} < t < \frac{l+1}{2}, l = 0, 1, \cdots, n \\ \sum_{k=0}^{n+1} a_k = A(1) = 0, \ \frac{n+1}{2} < t \end{cases} \quad (33)$$

is a locally supported piecewise constant function that can be easily generated by electric current switches.

The system S excited by the piecewise constant function

$$u(t) = \int_{-\infty}^{t} v(\tau)d\tau = \sum_{l=0}^{n} c_l \int_{-\infty}^{t-l/2} a(\tau)d\tau \quad (34)$$

shapes the pulse

$$p(t) = S(u)(t) \quad (35)$$

which is locally supported as

$$p(t) = 0, \quad t \notin (0, n) \quad (36)$$

and has the relationship

$$p(t) = \sum_{l=0}^{n} c_l \int_{-\infty}^{t-l/2} \alpha(\tau)d\tau = \int_{-\infty}^{t} q(\tau)d\tau. \quad (37)$$

Besides the simple and practical system (35) to shape $p(t)$ from the piecewise constant seed $u(t)$, the pulse $p(t)$ has the remarkable property

$$\int_{-\infty}^{\infty} \frac{d^2}{dt^2}p(t)\ p(t-k)dt = -\int_{-\infty}^{\infty} q(t)q(t-k)dt = \begin{cases} -1, & k = 0 \\ 0, & k = \pm 1, \pm 2, \cdots \end{cases} \quad (38)$$

which follows from (17), (18), (36), (37) and the partial integration formula. This property gives the foundation to transmission and detection of the pulse $p(t)$ in the impulse radio communications.

Given data bits $\{w_l\}$, we transmit the waveform

$$w(t) = S\left(\sum_{l=-\infty}^{\infty} w_l u(t-l)\right) = \sum_{l=-\infty}^{\infty} w_l p(t-l) \quad (39)$$

as illustrated in Fig. 2. Since a good broadband antenna is well approximated [6] by $\frac{d}{dt}$, the transmitted signal $w(t)$ is differentiated once by the transmitter antenna to be the radio signal

$$\frac{d}{dt}w(t) = \sum_{l=-\infty}^{\infty} w_l \frac{d}{dt}p(t-l) \quad (40)$$

and again by the receiver antenna to arrive at the receiver as

$$\frac{d^2}{dt^2}w(t) = \sum_{l=-\infty}^{\infty} w_l \frac{d^2}{dt^2}p(t-l). \quad (41)$$

Correlating the received signal $\frac{d^2}{dt^2}w(t)$ with the template pulse $p(t-k)$, which is the same as the transmission pulse, for its duration $(k, k+n)$, we have the bit w_k recovered by

$$\begin{aligned} \int_k^{k+n} \frac{d^2}{dt^2}w(t)\ p(t-k)dt &= \int_{-\infty}^{\infty} \frac{d^2}{dt^2}w(t)\ p(t-k)dt \\ &= \sum_{l=-\infty}^{\infty} w_l \int_{-\infty}^{\infty} \frac{d^2}{dt^2}p(t-l)\ p(t-k)dt \\ &= -w_k \end{aligned} \quad (42)$$

because of the property (38).

It should be noted that, because of (38), the detection formula (42) virtually performs the least-squares approximation of the radio waveform $\frac{d}{dt}w(t)$ by $\frac{d}{dt}p(t-k) = q(t-k)$ to detect w_k. Additive noises superimposed on $\frac{d}{dt}w(t)$ will then be most suppressed in the sense of least-squares estimation.

An example pulse associated with the transfer function

$$G(s) = \frac{1}{(s+18)(s+11.1i+10^{-13})(s-11.1i+10^{-13})} \quad (43)$$

and its derivatives are plotted in Fig. 3. The correlation in Fig. 4 becomes 1 and 0 at the origin and at the other integers, respectively, to verify (38). The power spectral density of the radio pulse $\frac{d}{dt}p(t) = q(t)$ is plotted in Fig. 5 along with the spectral mask (plotted by the boxy line) for the indoor ultra-wideband communications systems [1] imposed by the US Federal Communications Commission

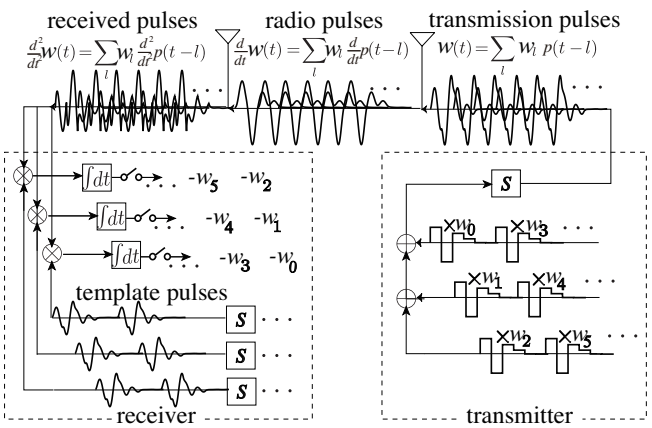
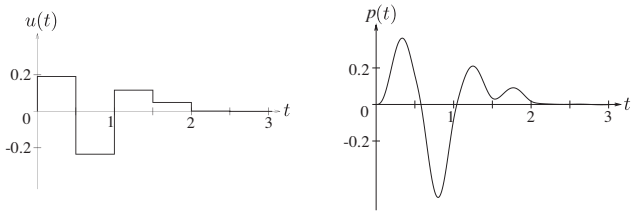
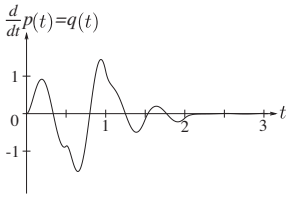


Figure 2: Schematic diagram of the transceiver.

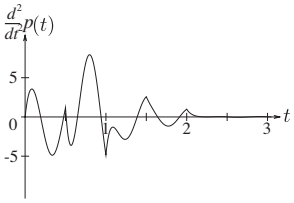


(a) pulse seed $u(t)$

(b) transmission pulse $p(t)$



(c) radio pulse $\frac{d}{dt}p(t)$



(d) received pulse $\frac{d^2}{dt^2}p(t)$

Figure 3: Pulses for impulse radio.

as the upper bound which no practical pulses are allowed to exceed. The frequency axis of the mask is scaled down by 6 GHz for the purpose of comparison, or equivalently, the pulse repetition rate is assumed to be 6 G pulses per second, which is much faster than the 1.32G pulses per second of the high speed direct sequence ultra-wideband protocol discussed in the IEEE 802.15.3a standard.

The fast transmission is possible because the pulses are orthogonal even though they are densely overlapping. But dense pulses are prone to interfere with one another in the situation that several reflected pulses arrive with various delays. Multipath compensation by digital filtering is crucial in order to effectively exploit the dense pulses we obtained. Transmitting a sounder pulse and digitizing the observed correlations, we have the end-to-end impulse response of the multipath channel. Digital filtering by an FIR approximation of the inverse impulse response will work as a kind of rake receiver. This compensation requires an analog-to-digital converter and a digital filter that work at the pulse rate and thus costs more hardware. But this cost should be justified since all the pulse-based systems cannot be faster without having denser pulses in the first place. A detailed analysis of the multipath effects, channel modeling error, and pulse synchronization is available in [3].

We may ignore the multipath effects and channel modeling error in the extreme situation that antennas are inductively coupled at a very short distance less than one inch. TransferJet technology has been working in the same situation at the maximum transmission rate of 560Mbps since 2008. A faster system will hopefully be the first application of the dense pulses obtained in this paper.

4. Conclusions

Inspired by the M-shaped orthogonal pulse, we derived a procedure to construct an exponential spline pulse with the knot interval $\frac{1}{2}$ that is locally supported, has its mean zero, and is orthogonal to its shifts by integers. An example pulse was obtained that will potentially enable an impulse radio communications system as fast as 6G pulses per second under the FCC regulation for the indoor ultra-wideband communications.

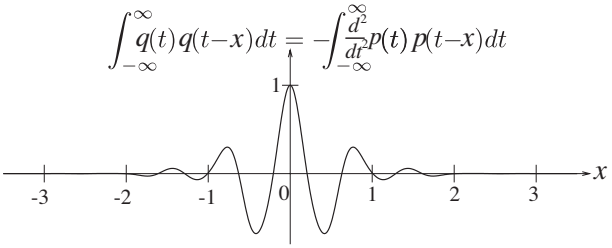


Figure 4: Correlation of the pulse.

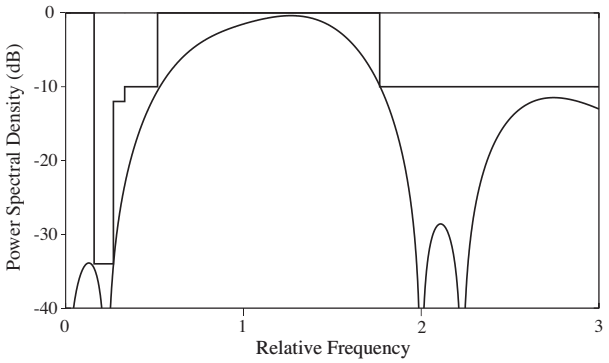


Figure 5: Power spectral density of the pulse and the FCC spectral mask.

5. Acknowledgment

This work was partially supported by JSPS grant-in-aid No. 17560357.

References:

[1] *Revision of Part 15 the Commission's rule regarding ultra-wideband transmission systems*. ET Docket No.98-153, Federal Communications Commission, Washington, D.C., 2002.

[2] A. J. Jerri. *Wavelets – Detailed Treatment with Applications*. Exercises of Chapter 3. Sampling Publishing, Potsdam, NY, to appear in 2009.

[3] M. Kamada S. Özlem and H.Habuchi. Construction of orthogonal overlapping pulses for impulse radio communications. *IEICE Transactions on Fundamentals*, E91-A(11):3121–3129, Nov. 2008.

[4] M. Unser and T. Blu. Cardinal exponential splines: Part I—Theory and filtering algorithms. *IEEE Transactions on Signal Processing*, 53(4):1425–1438, April 2005.

[5] M. Unser. Cardinal exponential splines: Part II—Think analog, act digital. *IEEE Transactions on Signal Processing*, 53(4):1439–1449, April 2005.

[6] M. Z. Win and R. A. Scholtz. Impulse radio: how it works. *IEEE Commun. Lett.*, 2(2):36–38, 1988.

Special session on

Mathematical Aspects of Compressed Sensing

Chair: Holger RAUHUT

A short note on non-convex compressed sensing

Rayan Saab [(1)] and Özgür Yılmaz[(2)]

(1) Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, B.C. Canada V6T 1Z4.
(2) Department of Mathematics, University of British Columbia, Vancouver, B.C. Canada V6T 1Z2.
rayans@ece.ubc.ca, oyilmaz@math.ubc.ca

Abstract:

In this note, we summarize the results we recently proved in [14] on the theoretical performance guarantees of the decoders Δ_p. These decoders rely on ℓ^p minimization with $p \in (0,1)$ to recover estimates of sparse and compressible signals from incomplete and inaccurate measurements. Our guarantees generalize the results of [2] and [16] about decoding by ℓ_p minimization with $p = 1$, to the setting where $p \in (0,1)$ and are obtained under weaker sufficient conditions. We also present novel extensions of our results in [14] that follow from the recent work of DeVore et al. in [8]. Finally, we show some insightful numerical experiments displaying the trade-off in the choice of $p \in (0,1]$ depending on certain properties of the input signal.

1. Introduction

Let Σ_S^N be the set of all S-sparse vectors,

$$\Sigma_S^N := \{x \in \mathbb{R}^N : \quad \#\text{supp}(x) \leq S\},$$

and define, qualitatively, compressible vectors as vectors that can be "well approximated" in Σ_S^N. For $p > 0$, let $\sigma_S(x)_{\ell^p}$ denote the best S-term approximation error of x in ℓ^p (quasi-)norm, i.e.,

$$\sigma_S(x)_{\ell^p} := \min_{v \in \Sigma_S^N} \|x - v\|_p.$$

We are interested in recovering x from its possibly noisy "encoding"

$$b = Ax + e, \tag{1}$$

where A is an $M \times N$ matrix with $M < N$. Equivalently, we seek accurate, stable, and "implementable" decoders $\Delta : \mathbb{R}^M \mapsto \mathbb{R}^N$ such that $\|\Delta(Ax + e) - x\|$ scales well with the noise level $\|e\|$, and is small whenever x is compressible.

In general, the problem of constructing decoders with such properties is non-trivial (even if $e = 0$) as A is overcomplete. However, if $A \in \mathbb{R}^{M \times N}$ is in general position, it can be shown that there is a decoder Δ_0 which satisfies $\Delta_0(Ax) = x$ for all $x \in \Sigma_S^N$ whenever $S < M/2$ [10]. This Δ_0 can be explicitly computed via the optimization problem

$$\Delta_0(b) := \arg\min_y \|y\|_0 \text{ subject to } b = Ay. \tag{2}$$

Unfortunately, (2) is combinatorial in nature, thus its complexity grows extremely quickly as N becomes much larger than M. Naturally, one then seeks to replace (2) with a more tractable optimization problem.

1.1 Decoding by ℓ^p minimization

Define the decoders

$$\Delta_p^\epsilon(b) = \arg\min_x \|x\|_p \text{ subject to } \|Ax - b\|_2 \leq \epsilon, \tag{3}$$

and

$$\Delta_p(b) = \arg\min \|x\|_p \text{ subject to } Ax = b, \tag{4}$$

with $0 < p \leq 1$. [2, 4, 9, 10, 15], that in the noise-free setting Δ_1 recovers x exactly if x is sufficiently sparse and if A has certain properties. Furthermore, one has error guarantees even when x is not "exactly" sparse and when the encoding is noisy, e.g., [2, 9].

In this note we focus on Δ_p and Δ_p^ϵ with $0 < p < 1$. Early work by Gribonval and co-authors (e.g. [12, 13]) presents sufficient conditions for having $\Delta_p(b) = \Delta_1(b) = x$ and stability conditions to deal with noisy encoding. However, these conditions are pessimistic in the sense that they generally guarantee recovery of only very sparse vectors.

Recently, Chartrand [5] showed that in the noise-free setting, a sufficiently sparse signal can be recovered perfectly with Δ_p, where $p \in (0,1)$, under less restrictive requirements than those needed to guarantee perfect recovery with Δ_1. Moreover, in [6], Staneva and Chartrand showed that if A is an $M \times N$ Gaussian matrix, recovery of x in Σ_S^N is guaranteed provided $M > C_1(p)S + pC_2(p)S\log(N/K)$. In other words, the dependence on N of the required number of measurements M (that guarantees perfect recovery for all $x \in \Sigma_S^N$) disappears as p approaches 0, unlike the case with $p = 1$. These results motivate a more detailed study of the stability and robustness properties of the decoders Δ_p.

In the remainder of the note, we summarize our recent results in [14] concerning the theoretical properties of Δ_p and Δ_p^ϵ. In addition, we present some extensions of our results on the instance optimality in probability of Δ_p when the entries of A are drawn from any sub-Gaussian distribution. Finally, we present numerical results suggesting scenarios where using Δ_p, $p \in (0,1)$, is better than using Δ_1.

2. Main Results

We begin with the relevant notation. Let δ_S, the S-restricted isometry constants of A (see, e.g., [2]), be the smallest constants satisfying

$$(1-\delta_S)\|c\|_2^2 \leq \|Ac\|_2^2 \leq (1+\delta_S)\|c\|_2^2$$

for every $c \in \Sigma_S^N$. We say that a matrix satisfies RIP(S,δ) if $\delta_S < \delta$. It has been shown that if A is an $M \times N$ matrix the columns of which are i.i.d. random vectors with any sub-Gaussian distribution, then A satisfies RIP (S,δ) with $S \leq c_1 M/log(N/M)$, $\delta < 1$ with probability $> 1-2e^{-c_2 M}$ (see, e.g., [1], [3]). Following the notation of [16], we say that a decoder Δ is (q,p) *instance optimal* if

$$\|\Delta(Ax)-x\|_q \leq C\sigma_S(x)_{\ell^p}/S^{1/p-1/q} \qquad (5)$$

holds for all $x \in \mathbb{R}^N$. Moreover, a decoder Δ is said to be (q,p) *instance optimal in probability* if (5) holds for any x with high probability on the draw of A. Note that the stability results of Candès et al. [2] imply (2,1) instance optimality of the decoder Δ_1, while the results of Wojtaszczyk in [16] show that Δ_1 is (2,2) instance optimal in probability if the entries of A are drawn from a Gaussian distribution or if its columns are drawn uniformly from the sphere.

2.1 Decoding with Δ_p: stability and robustness

We consider the scenario where x is arbitrary and $\sigma_S(x)_{\ell^p}$ is its best S-term approximation error measured in ℓ^p (qausi)-norm. In particular, we are interested in controlling the error $\|\Delta_p^\epsilon(b)-x\|_2^p$.

Theorem 1 *Let $p \in (0,1]$ and let x be arbitrary. Suppose that*

$$\delta_{kS} + k^{\frac{2}{p}-1}\delta_{(k+1)S} < k^{\frac{2}{p}-1} - 1, \qquad (6)$$

for some $k > 1$, $kS \in \mathbb{Z}^+$. Let $b = Ax + e$ where $\|e\|_2 \leq \epsilon$. Then $\Delta_p^\epsilon(b)$ satisfies

$$\|\Delta_p^\epsilon(b)-x\|_2^p \leq C^{(1)}\epsilon^p + C^{(2)}\frac{\sigma_S(x)_{\ell^p}^p}{S^{1-p/2}}, \qquad (7)$$

where $C^{(1)}$ and $C^{(2)}$ are given in [14].

Remark 2 *This is a straightforward generalization of the results of [2] regarding the performance of Δ_1. In fact, by setting $p = 1$ in the above theorem, we obtain the main theorem of [2], with precisely the same constants.*

Remark 3 *Using $\epsilon = 0$ in the above theorem, we find that the decoder Δ_p is $(2,p)$ instance optimal. Similarly, assuming $x \in \Sigma_S^N$ (hence $\sigma_S(x)_{\ell^p} = 0$), we see that Δ_p^ϵ can stably recover sparse signals.*

We can also compare S_p, the sparsity of vectors that are guaranteed to be recovered with Δ_p and S_1, the sparsity of vectors that are guaranteed to be recovered with Δ_1. This helps illustrate the possible benefits of using Δ_p over using Δ_1 in recovering sparse signals.

Corollary 4 (relationship between S_1 and S_p) *Suppose for some k and S_1, $\delta_{(k+1)S_1} < \frac{k-1}{k+1}$. Then Δ_1 recovers S_1-sparse vectors and Δ_p recovers S_p-sparse vectors with*

$$S_p \geq \left\lfloor \frac{k+1}{k^{p/(2-p)}+1} S_1 \right\rfloor.$$

2.2 Instance optimality in probability of Δ_p

In [7], it was shown that no decoder, $\Delta : \mathbb{R}^M \mapsto \mathbb{R}^N$, is $(2,2)$ instance optimal unless $M \sim N$. In this section, we show that Δ_p is $(2,2)$ instance optimal in probability. Our approach is similar to that of [16], which we summarize now. Denoting by B_q^K the unit ball of ℓ^q in K dimensions, a matrix A is said to possess the $LQ_1(\alpha)$ property if and only if

$$A(B_1^N) \supset \alpha B_2^M.$$

In [16], Wojtaszczyk shows that random Gaussian matrices of size $M \times N$, as well as matrices whose columns are drawn uniformly from the sphere posses the $LQ_1(\alpha)$ property, $\alpha = \mu\sqrt{\frac{\log(N/M)}{M}}$ with high probability. Here $\mu < 1/\sqrt{2}$ is a constant. Noting that such matrices also satisfy $RIP((k+1)S,\delta)$ with $S < c\frac{M}{log(N/M)}$ with high probability, Wojtaszczyk proves that Δ_1, with these matrices, is (2,2) instance optimal in probability. Our proof of the analogous result for Δ_p, $p \in (0,1)$, relies on the non-trivial generalization of the LQ_1 property to an $LQ_p(\alpha)$ property with $\alpha = 1/C_p\left(\mu^2\frac{\log(N/M)}{M}\right)^{(1/p-1/2)}$. Specifically, we say that a matrix A satisfies $LQ_p(\alpha)$ if and only if

$$A(B_p^N) \supset \alpha B_2^M.$$

Below, we will use A_ω to denote matrices whose entries are drawn from a zero mean, normalized column variance Gaussian distribution and $\tilde{A}_\omega$ to denote matrices drawn uniformly from the sphere. The following lemma states that the matrices A_ω and $\tilde{A}_\omega$ satisfy the LQ_p property with high probability.

Lemma 5 *$\tilde{A}_\omega$ and A_ω satisfy the $LQ_p(\alpha)$ property with $\alpha = 1/C_p\left(\mu^2\frac{\log(N/M)}{M}\right)^{1/p-1/2}$ with probability $\geq 1-e^{-cM}$ on the draw of the matrix. Here, C_p is a constant that depends only on p, $\mu < 1/\sqrt{2}$ is a constant, and c is a constant that depends on μ.*

Proving Lemma 5 is non-trivial and requires a result by [11], relating the distances of p-convex bodies to their convex hulls. On the other hand, this lemma provides the machinery needed to prove the following theorem, which extends an analogous result of Wojtaszczyk [16].

Theorem 6 *Let $A_\omega \in \mathbb{R}^{M\times N}$, $\omega \in \Omega$, be a random matrix with entries drawn independently from a zero-mean, normalized column variance Gaussian distribution, and let (Ω, P) be the associated probability space. There exists constants $c_1, c_2, c_3 > 0$ such that for all $S \leq c_1 M/\log(N/M)$, the following are true.*

(i) $\exists\Omega_1$, with $P(\Omega_1) \geq 1-e^{-c_2 M}$, such that $\forall x \in \mathbb{R}^N$, $\forall e \in \mathbb{R}^M$ and $\forall\omega \in \Omega_1$

$$\|\Delta_p(A_\omega(x)+e)-x\|_2 \leq C(\|e\|_2 + \frac{\sigma_S(x)_{\ell^p}}{S^{1/p-1/2}}), \quad (8)$$

(ii) $\forall x \in \mathbb{R}^N$, $\exists \Omega_x$, *with* $P(\Omega_x) \geq 1 - e^{-c_3 M}$, *such that* $\forall e \in \mathbb{R}^M$ *and* $\forall \omega \in \Omega_x$

$$\|\Delta_p(A_\omega(x)+e)-x\|_2 \leq C\left(\|e\|_2 + \sigma_S(x)_{\ell^2}\right). \quad (9)$$

The statement also holds for $\widetilde{A}_\omega$.

Note that the constants above (both denoted by C) rely on the parameters of the particular LQ_p and RIP properties that the matrix satisfies, and are omitted for ease of exposition. For the proofs of Lemma 5 and Theorem 6 see [14]. Finally, we present the following extension of Theorem 6.

Proposition 7 *The conclusions of Theorem 6 also hold when the entries of* A *are i.i.d., drawn from a sub-Gaussian distribution.*

Our proof of the above proposition, which we omit here, relies on the recent work of [8] where the $LQ_1(\alpha)$ property was modified, allowing the authors to show the (2,2) instance optimality of Δ_1 when the entries of the matrix A are drawn from any sub-Guassian distribution.

3. Numerical Experiments

In this section, we present some numerical experiments to highlight important aspects of sparse recovery using Δ_p, $0 < p \leq 1$. First, we are interested in the sufficient conditions under which decoding with Δ_p can guarantee perfect recovery of signals in Σ_S^N for different values of p and S. Our goal is to show empirically that with smaller values of $p \in (0,1)$, Δ_p allows recovery of less sparse signals than would have been possible with larger values of p, as Theorem 1 predicts.

To that end, we generate a 100×300 matrix whose columns are drawn from a Gaussian distribution and estimate its RIP constants δ_S via Monte Carlo (MC) simulations. Under the assumption that the estimated constants are the correct ones (while in fact they are only lower bounds), Figure 1(a) shows the regions where (6) guarantees recovery for different (S,p)-pairs. On the other hand, Figure 1(b) shows the empirical recovery rates using the same matrix with fifty different instances of $x \in \Sigma_S^N$, and decoding by Δ_p, where we choose the non-zero coefficients of x randomly from the Gaussian distribution. Here, we compute $\Delta_p(Ax)$, as a solution to the ℓ^p optimization problem of (4) by using a projected gradient algorithm on a smoothed version of $\|x\|_p^p$, namely $\sum_i (x_i^2 + \epsilon^2)^{p/2}$, where the solution to each subproblem, starting with a large ϵ is used as an initial estimate for the next subproblem with a smaller ϵ. Note that this approach is similar to the one described in [5]. Clearly, the empirical results show that Δ_p is successful in a wider range of scenarios than those predicted by Theorem 1. This can be attributed to the fact that the conditions presented in this paper are only sufficient. Moreover, what is observed in practice is not necessarily a manifestation of uniform recovery. Rather, the practical results could be interpreted as success of Δ_p with high probability on either x or A.

In our second set of experiments, we wish to observe the instance optimality of Δ_p, i.e., the linear growth of the

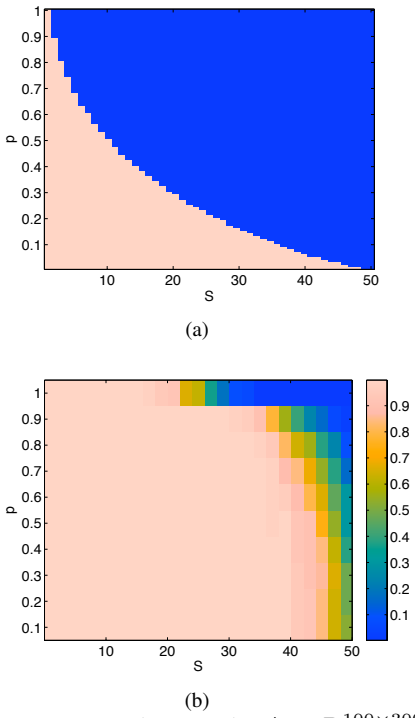


Figure 1: For a Gaussian matrix $A \in \mathbb{R}^{100\times 300}$, whose δ_S values are estimated via MC simulations, we generate the theoretical (a) and practical (b) phase-diagrams for reconstruction via ℓ^p minimization. The lighter shading indicates higher recoverability rates. .

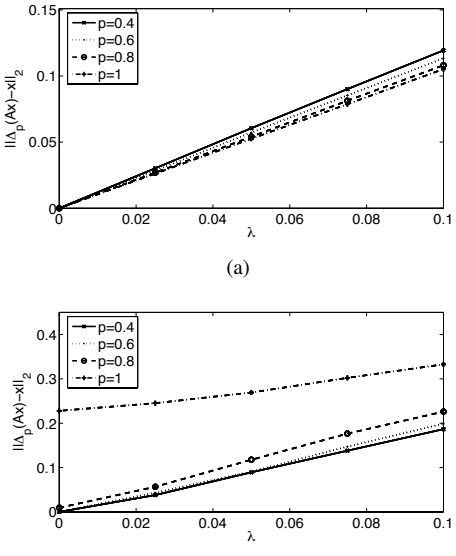


Figure 2: Reconstruction error with compressible signals, $S = 5$ (a), $S = 35$ (b). Observe the almost linear growth of the error for different values of p, highlighting the instance optimality in probability of the decoders.

ℓ^2 reconstruction error $\|\Delta_p(Ax) - x\|_2$, as a function of $\sigma_S(x)_{\ell^2}$. To that end, we generate scenarios that allude to the conclusions of Theorem 6. We generate a signal composed of $x_T \in \Sigma_S^{300}$, supported on an index set T, and a signal z_{T^c} supported on $T^c = \{1,2,...,300\}\backslash T$, where all the coefficients are drawn from the Gaussian distribution and $\|x_T\|_2 = \|z_{T^c}\|_2 = 1$. We then set $x_\lambda = x_T + \lambda z_{T^c}$ with increasing values of λ starting from 0, i.e., x_λ be-

comes less compressible as λ increases, and T is the "effective support" of x_λ. Next, we choose our measurement matrix $A \in \mathbb{R}^{100\times 300}$ by drawing its columns uniformly from the sphere. For each value of λ we measure the reconstruction error $\|\Delta_p(Ax_\lambda) - x_\lambda\|_2$, and we repeat the process 50 times while randomizing the index set T but preserving the coefficient values. We report the averaged results for different values of p with $S = 5$ in Figure 2(a) and $S = 35$ in Figure 2(b). Note that when $S = 5$, Δ_1 provides the best performance, and the performance of Δ_p degrades monotonically as p decreases. On the other hand, when $S = 35$, Δ_p with $p = 0.4$ provides the best performance and the performance degrades as p increases.

We investigate this observation further by examining the performance as a function of $S \in \{5, 10, ..., 35\}$. In Figure 3, we plot the value of an "empirical effective constant" which we calculate as the maximum of $\|\Delta_p(Ax_\lambda) - x_\lambda\|_2/\lambda$ as $\lambda > 0$ varies. This constant acts as a surrogate for C in (9) assuming that such a constant exists and that $\sigma_S(x)_{\ell^2} = \|\lambda z_{T^c}\|_2 = \lambda$. The behavior gradually changes from favoring $p = 1$ when S, the size of the effective support of x_λ, is small to favoring $p = 0.4$ as S increases.

A closer look at the explicit value of the constant in Theorem 6 sheds some light on this behavior. Below, we use the notation of [14]. The constant C in (9) behaves like $(2C^{(2)})^{1/p}/\gamma_p$ (where $C^{(2)}$ and γ_p are explicitly given in [14]). Specifically, $1/\gamma_p$ depends only on the matrix A and increases exponentially as p decreases, while $C^{(2)}$, the constant in Theorem 1, depends on p, as well as k and $\delta_{(k+1)S}$ (where $k > 1$ is a free parameter). When S is relatively small, the associated RIP constants remain small, which consequently implies that $[C^{(2)}]^{1/p}$ remains small provided p is isolated away from 0. In this case, the behavior of C is determined by that of γ_p, i.e., C is smallest when $p = 1$. On the other hand, when S is large, $[C^{(2)}]^{1/p}$ grows as p approaches 1 (this is a manifestation of the more restrictive RIP requirements for larger p as stated in (6)). This increase seems to be dominating the behavior of C, thus for larger S we get better effective constants with smaller p. Such a heuristic could be an interpretation of the behavior we observe in Figure 3. For a rigorous quantitative analysis, one needs to identify the s-restricted isometry constants of the matrix A for every s. Such a treatment is beyond the scope of this note.

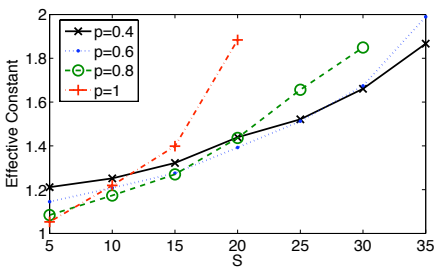


Figure 3: The empirical effective constant as a function of S for different values of p. Note the gradual change favoring $p = 1$ when S is small to $p = 0.4$ as S increases.

References:

[1] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A Simple Proof of the Restricted Isometry Property for Random Matrices. *Constructive Approximation*, 2008.

[2] E. J. Candès, J. Romberg, and T. Tao. Signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1207–1223, 2005.

[3] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51(12):489–509, 2005.

[4] E. J. Candès and T. Tao. Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theory*, 52(12):5406–5425, 2006.

[5] R. Chartrand. Exact reconstructions of sparse signals via nonconvex minimization. *IEEE Signal Process. Lett.*, 14(10):707–710, 2007.

[6] R. Chartrand and V. Staneva. Restricted isometry properties and nonconvex compressive sensing. *Inverse Problems*, 24(035020), 2008.

[7] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *Journal of the American Mathematical Society (to appear)*, 2008.

[8] R. DeVore, G. Petrova, and P. Wojtaszczyk. Instance-optimality in probability with an ℓ_1-minimization decoder. *preprint*, 2008.

[9] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[10] D. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization. *Proc. Natl. Acad. Sci. USA*, 100(5):2197–2202, 2003.

[11] Y. Gordon and N.J. Kalton. Local structure theory for quasi-normed spaces. *Bull. Sci. Math.*, 118:441–453, 1994.

[12] R. Gribonval, R. M. Figueras i Ventura, and P. Vandergheynst. A simple test to check the optimality of sparse signal approximations. *EURASIP Signal Processing, special issue on Sparse Approximations in Signal and Image Processing*, 86(3):496–510, 2006.

[13] R. Gribonval and M. Nielsen. Highly sparse representations from dictionaries are unique and independent of the sparseness measure. *Appl. Comput. Harm. Anal.*, 22(3):335–355, May 2007.

[14] R. Saab and O. Yilmaz. Sparse recovery by nonconvex optimization – instance optimality. *CoRR*, abs/0809.0745, 2008.

[15] J.A. Tropp. Recovery of short, complex linear combinations via l^1 minimization. *IEEE Transactions on Information Theory*, 51(4):1568–1570, April 2005.

[16] P. Wojtaszczyk. Stability and instance optimality for gaussian measurements in compressed sensing. *Preprint*, 2008.

Orthogonal Matching Pursuit with random dictionaries

P. Bechler, and P. Wojtaszczyk

Institut of Applied Mathematics, University of Warsaw
P.Bechler@mimuw.edu.pl, wojtaszczyk@mimuw.edu.pl

Abstract:

In this paper we investigatet the efficiency of the Orthogonal Matching Pursuit for random dictionaries. We concentrate on dictionaries satisfying Restricted Isometry Property. We introduce a stronger Homogenous Restricted Isometry Property which is satisfied with overwhelming probability for random dictionaries used in compressed sensing. We also present and discuss some open problems about OMP.

1. Introduction

In this paper we investigate the efficiency of the Orthogonal Matchin$T = U\sqrt{T^*T}$g Pursuit for random dictionaries. Orthogonal Matching Pursuit is a well known greedy algorithm widely used in approximation theory, statistical estimations and compressed sensing (for the review of greedy algorithms see [6]). One of its main features is that it can be applied for arbitrary dictionary. However the efficiency of the algorithm depend very strongly on properties of the dictionary. We work in the context of a Hilbert space $\mathcal{H}$ (which may be assumed to be finite dimensional). The dictionary is a subset $\mathcal{D} \subset \mathcal{H}$ such that $\overline{\operatorname{span} \mathcal{D}} = \mathcal{H}$. We usually assume that $\|x\|$ is close to 1 for $x \in \mathcal{D}$. Generally it is assumed that $\|x\| = 1$ for $x \in \mathcal{D}$ (see e.g. [6]). However for random dictionaries it is very rarely satisfied. On the other hand for such dictionary $\|x\|$ is close to 1 with great probability.

The Orthogonal Matching Pursuit algorithm with respect to the dictionary $\mathcal{D}$ obtains iteratively a sequence $\operatorname{OMP}_n f \in \mathcal{H}$ of approximants of a given element $f \in \mathcal{H}$ and a sequence $d_1, \dots, d_n \in \mathcal{D}$ in the following way:

- Define $\operatorname{OMP}_0 f = 0$.
- Given $\operatorname{OMP}_{n-1} f$ and $d_1, \dots, d_{n-1} \in \mathcal{D}$ choose $d_n \in \mathcal{D}$ such that

  $$|\langle f - \operatorname*{OMP}_{n-1} f, d_n \rangle| = \sup \left\{ |\langle f - \operatorname*{OMP}_{n-1} f, d \rangle \ : d \in \mathcal{D} \right\}$$

  and define $\operatorname{OMP}_n f$ as the orthogonal projection of f onto $\operatorname{span}\{d_1, \dots, d_n\}$.

Generally we will write $f - \operatorname{OMP}_s f := f_s$.

The standard measure of approximation power of a dictionary is the error of the best m–term approximation. We define the set of m sparse vectors (with respect to the dictionary $\mathcal{D}$) as

$$\Sigma_m^{\mathcal{D}} = \Sigma_m = \{ \sum_{j=1}^{m} a_j d_j \ : \ \{d_j\}_{j=1}^m \subset \mathcal{D} \}. \tag{1}$$

For a given $f \in \mathcal{H}$ we define its best error of m–term approximation as

$$\sigma_m(f) = \inf\{ \|f - z\| \ : \ z \in \Sigma_m \}. \tag{2}$$

Clearly we always have $\sigma_m(f) \leq \|f - \operatorname{OMP}_m(f)\| = \|f_m\|$.

Obviously when our dictionary is an orthonormal basis then $\sigma_m(f) = \|f - \operatorname{OMP}_m(f)\|$ for each $f \in \mathcal{H}$. Unfortunately this is the only case when it is so. The fundamental, and still largely unanswered question is how close $\operatorname{OMP}_m(f)$ can get to this optimal rate of approximation given by $\sigma_m(f)$. It is to be expected that the answer to the above question must depend on some extra properties of the dictionary.

2. Dictionaries

One of the commonly used quantitative parameters of the dictionary is its mutual coherence. It is defined as

$$\eta = \sup_{d_1 \neq d_2 \in \mathcal{D}} |\langle d_1, d_2 \rangle|. \tag{3}$$

Recently, especially in the context of compressed sensing, a restricted isometry property (RIP for short) became very useful. Let us recall the following well known definition (c.f. [1, 2]).

Definition 1 *The dictionary $\mathbf{\Phi} = \{\phi_j\}_{j=1}^N$ has $\operatorname{RIP}(k, \delta)$, $0 < \delta < 1$ if for any set $T \subset \{1, \dots, N\}$ with $\#T = k$ and any sequence of numbers x_j we have*

$$(1-\delta)\sqrt{\sum_{j \in T} |x_j|^2} \leq \| \sum_{j \in T} x_j \phi_j \| \leq (1+\delta)\sqrt{\sum_{j \in T} |x_j|^2}. \tag{4}$$

There are some easy relations between those notions. If the dictionary $\mathcal{D}$ has mutual coherence η then it satisfies $\operatorname{RIP}(k, 1-\eta)$ for $k < \eta^{-1}$. On the other hand if $\mathcal{D}$ satisfies $\operatorname{RIP}(k, \delta)$ then it has mutual coherence $\sim \delta$.

Usually dictionaries with RIP are exhibited as random dictionaries. To be more precise we define a dictionary in $\mathbb{R}^n$

as $\mathbf{\Phi}(\omega) = \{\phi_j\}_{j=1}^N$ where $\phi_j = (\gamma_{j,1}, \ldots, \gamma_{j,n})$ and $\gamma_{j,i}$ are idependent copies of a fixed subgaussian random variable normalised so that $\mathbb{E}\|\phi_k\|^2 = 1$.
In this context it is known (see e.g. [1]) that for a fixed $0 < \delta < 1$ there exists $c > 0$ such that the dictionary $\mathbf{\Phi}(\omega)$ with overwhelming probability satisfies $\mathrm{RIP}(k, \delta)$ with $k = \lfloor cn/\log N \rfloor$. On the other hand it is also known that such a dictionary with overwhelming probability has mutual coherence of order $k^{-1/2}$. It is clear that when we have two events each of them happening with big probability then they happen simultanously with big probability. This leads to the following definition:

Definition 2 *The dictionary $\mathbf{\Phi}$ has homogenous restricted isometry property* $\mathrm{HRIP}(k, \delta)$, $0 < \delta < 1$ *if for any $l \leq k$ it satisfies* $\mathrm{RIP}(l, \delta\sqrt{l/k})$.

Following standard reasoning we obtain

Theorem 1 *Suppose that integers n, N and numbers $0 < \delta < 1$ and $a > 0$ are given and suppose that the random dictionary $\mathbf{\Phi}(\omega) = \{\phi_1, \ldots, \phi_N\} \subset \mathbb{R}^n$ is as described above. Then there exist $c, c_1 > 0$ which depend only on the subgaussian distribution involved, δ and a such that dictionary $\mathbf{\Phi}(\omega)$ satisfies* $\mathrm{HRIP}(k, \delta)$ *for $k = \lfloor c_1 n/\log N \rfloor$ with probability $\geq 1 - 3N^{-a}$*

Basically this tells us that unless we are very unlucky a randomly chosen dictionary satisfies HRIP, which is clearly stronger property than RIP. We believe that HRIP is a useful property. Theorem 4 is some indication of this.

3. Main Result

Now we want to present a result on the approximation power of OMP for dictionaries satisfying RIP. For dictionaries with incoherence analogous results were obtained by D. Donoho, M. Elad and N.V. Temlyakov [3]. If we are interested in random dictionaries results from [3] require $S \leq \sqrt{n/\log N}$ while ours apply for the full range $S \leq n/\log N$.

Theorem 2 *There exist constants C and c depending only on $\epsilon > 0$ such that for the dictionary $\mathbf{\Phi}$ satisfying* $\mathrm{RIP}(2K, \epsilon)$ *and for $0 \leq k \leq S \leq K$ we have*

$$\|f_S\|^2 \leq C\|f_k\| \left(\sigma_{S-k}(f_k) + A\epsilon\|f_k\|\right). \tag{5}$$

with $A = c(1 + \log K)$).

Note that in particular seting $k = 0$ we get

$$\|f_S\|^2 \leq C\|f\|(\sigma_S(f) + A\epsilon\|f\|). \tag{6}$$

The proof of Theorem 2 is rather complicated. It uses a lot of geometry of Hilbert space, theory of Riesz bases and ideas from [3] and [5]. The main new technical tool is the following lemma on norm of matrices.

Lemma 1 *Let $0 < \epsilon < 1$ and let $A = [a_{i,j}]$ be an $n \times n$ upper triangular matrix such that for any $x \in \mathbb{R}^n$*

$$(1 - \epsilon)\|x\| \leq \|Ax\| \leq (1 + \epsilon)\|x\| \tag{7}$$

and $|a_{i,i}| \geq 1 - \epsilon$ for $i = 1, 2, \ldots, n$. Let $B = [b_{i,j}]$ be the off diagonal part of A i.e.

$$b_{i,j} = \begin{cases} a_{i,j} & \text{if } i < j \\ 0 & \text{if } j \leq i. \end{cases}$$

Then $\|B\| \leq 4\epsilon\lceil \log_2 n \rceil$.

The above inequailties (5) and (6) have some merit only if $\epsilon A < 1$. Generally one would like to avoid the presence of $\|f_k\|$ (or $\|f\|$) inside the brackets in (5), (6). The most desirable would be to have direct estimates of the form $\|f_s\| \leq C\sigma_s(f)$. Unfortunatelly in full generality such estimates are not true even when we replace the constant by a function of s.
Here is an appropriate example. Let $x = \frac{1}{\sqrt{n}} \sum_{j=1}^n e_j \in \mathbb{R}^{2n}$ so $\|x\| = 1$. Let us consider the dictionary consisting of vectors: $e_1, \ldots, e_n$, $\psi_j := \|e_j + \beta n^{-1/2} x\|^{-1}(e_j + \beta n^{-1/2} x)$ for $j = n+1, \ldots, n+s$ plus orthonormal vectors which are orthonormal to all those to make a basis in $\mathbb{R}^{2n}$. We take $\beta = \sqrt[4]{n}$ and $s = \lfloor \epsilon\sqrt{n} \rfloor$. Then the following are easy to check

- The mutual coherence is $\leq n^{-1/2}$.
- The Riesz constant of this basis is $\sqrt{\epsilon}$ so the dictionary has $\mathrm{RIP}(2n, \sqrt{\epsilon})$
- Orthogonal Matching Pursuit for vector x in first s iterations chooses vectors ψ_j and only later chooses vectors e_j.

Thus we see that $\sigma_n(x) = 0$ while $x - \mathrm{OMP}_k(x) \neq 0$ for $k = n + s - 1$.
For dictionaries with mutual coherence η J. Tropp [7], slightly improving estimate from [4], have proved

Theorem 3 *If the dictionary has mutual coherence η then*

$$\|f_m\| \leq 8\sqrt{m}\sigma_m(f) \text{ for } m < (3\eta)^{-1}. \tag{8}$$

Using this we obtain

Theorem 4 *Let the dictionary $\mathbf{\Phi}$ satisfies* $\mathrm{HRIP}(k, \delta)$. *Then for $m \leq c/\sqrt{k}$ we have*

$$\|f_{\lfloor m \log m \rfloor}\| \leq C\sigma_m(f). \tag{9}$$

Let us give a sketch of a proof which follows arguments from [3]. We start with $m \leq c'\sqrt{k}$ for which (8) holds. We set $m_l = m(2^l - 1)$ and we fix $K \sim k^{3/4}$. Using HRIP we get that dictionary $\mathbf{\Phi}$ satisfies $\mathrm{RIP}(2K, \epsilon)$ with $A\epsilon \leq \delta k^{-1/8} \leq \beta m^{-1/4}$. From Theorem 2 and (8) we get

$$\begin{aligned} \|f_{m_2}\|^2 &\leq C\|f_{m_1}\|(\sigma_m(f) + A\epsilon\|f_{m_1}\|) \\ &\leq C\|f_{m_1}\|(\sigma_m(f) + 8\beta m^{1/4}\sigma_m(f)) \\ &\leq 8Cm^{1/2}(1 + 8\beta m^{1/4})\sigma_m^2(f) \\ &\leq C' m^{3/4}\sigma_m^2(f). \end{aligned}$$

Thus we get $\|f_{m_2}\| \leq \sqrt{C'} m^{3/8}\sigma_m(f)$. Repeating this argument and carefully tracking constants we see that after at most $\mu \sim \log\log m$ steps we get the claim.

Analogous result from [3] uses only mutual coherence and in our case gives (9) for $m \leq c\sqrt[3]{k}$. The main drawback of Theorem 4 is the limitation on m. It is clear from the above sketch that this restriction is inherited from Theorem 3. It is very unlikely that (8) can be substantially improved using only mutual coherence. We believe however that for dictionaries with RIP or HRIP one can prove more. So let us state the following conjecture

Conjecture Assume that the dictionary satisfies $\mathrm{HRIP}(k,\delta)$. There exist constants C, c, α and β (possibly depending on δ) such that for every f and for $m\log^{\alpha} m \leq ck$ we have

$$\|f_{\lfloor m\log^{\alpha} m\rfloor}\| \leq Cm^{\beta}\sigma_m(f).$$

Let us note that it follows from Theorem 3 that there exists a function $\psi(k,\delta)$ and constants C and β such that if the dictionary satisfies $\mathrm{HRIP}(k,\delta)$ then for every $f \in \mathcal{H}$

$$\|f_m\| \leq Cm^{\beta}\sigma_m(f).$$

for $m \leq \psi(k,\delta)$. (Clearly Theorem 3 gives $\beta = 1/2$ and $\psi(k,\delta) \sim \sqrt{k}$). It would be interesting to know if ψ can grow significantly faster than $\sqrt{k}$.

References:

[1] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, *A simple proof of the restricted isometry property for random matrices*, Constr. Approx., **28** (2008), no. 3, 253–263.

[2] E. Candés, *The restricted isometry property and its implications for compressed sensing*, Compte Rendus de l'Academie des Sciences, Paris, Series I, **346**(2008), 589–592.

[3] D.Donoho, M.Elad, V.N.Temlyakov, *On Lebesgue-type inequalities for greedy approximation* J. Approx. Theory **147** (2007), no. 2, 185–195.

[4] A.Gilbert, S.Muthukrishnan, M.Strauss, *Approximation of functions over redundant dictionaries using coherence*, Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms (Baltimore, MD, 2003), 243–252, ACM, New York, 2003.

[5] S.Kwapień, A. Pełczyński, *The main triangle projection in matrix spaces and its applications* Studia Math. **34** (1970) 43–68.

[6] V.N.Temlyakov, *Greedy approximation*, Acta Numerica **17** (2008) 235–409

[7] J.Tropp, *Greed is good: Algorithmic results for sparse approximation*, IEEE Trans. Inform. Theory, **50** (2004), 2231-2242

Average Case Analysis of Multichannel Basis Pursuit

Holger Rauhut [(1)], Yonina C. Eldar [(2)]

(1) Hausdorff Center for Mathematics, and Institute for Numerical Simulation, University of Bonn
Endenicher Allee 62, 53115 Bonn, Germany.
(2) Department of Electrical Engineering, Technion, Israel Institute of Technology, Haifa, Israel 32000.
rauhut@hcm.uni-bonn.de, yonina@ee.technion.ac.il

Abstract:

We consider the recovery of jointly sparse multichannel signals from incomplete measurements using convex relaxation methods. Worst case analysis is not able to provide insights into why joint sparse recovery is superior to applying standard sparse reconstruction methods to each channel individually. Therefore, we analyze an average case by imposing a probability model on the measured signals. We show that under a very mild condition on the sparsity and on the dictionary characteristics, measured for example by the coherence, the probability of recovery failure decays exponentially in the number of channels. This demonstrates that most of the time, multichannel sparse recovery is indeed superior to single channel methods.

1. Introduction

Recovery of sparse signals from a small number of measurements is a fundamental problem in many different signal processing tasks such as image denoising [3], analog-to-digital conversion [21, 11], radar, compression, inpainting, and many more. The recent framework of compressed sensing (CS), founded in the works of Donoho [8] and Candes [3], studies acquisition methods as well as efficient computational algorithms that allow reconstruction of a sparse vector x from linear measurements $y = Ax$, where A is referred to as the measurement matrix. The key observation is that y can be relatively short, and still contain enough information to recover x.

Determining the sparsest vector x consistent with the data $y = Ax$ is generally an NP-hard problem [7]. To determine x in practice, a multitude of efficient algorithms have been proposed. The most extensively studied recovery method by now is the ℓ_1-minimization approach (Basis Pursuit). Greedy methods, such as simple thresholding [23] or orthogonal matching pursuit (OMP) [26], are faster in practice, but BP provides significantly better recovery guarantees [10, 22].

The BP principle as well as greedy approaches have been extended to the multichannel setup where the signal consists of several channels [29, 30, 15, 6, 5, 20, 12, 13, 18]. Here one assumes that each channel is sparse and in addition that the channels have a small common support set. In this situation the signals are called jointly sparse. A variety of theoretical recovery results have been established already in this setting. In [5] a recovery result was derived for a mixed ℓ_p/ℓ_1 program (multichannel BP) in which the objective is to minimize the sum of the ℓ_p-norms of the rows of the estimated matrix whose columns are the unknown vectors.

Recovery results for the more general problem of block-sparsity were developed in [13] based on the restricted isometry property (RIP), and in [12] based on mutual coherence. In practice, multichannel reconstruction techniques perform much better than recovering each channel individually. However, the theoretical equivalence results predict no performance gain. The reason is that these recovery results apply to all possible input signals, and are therefore worst-case measurements. Clearly, if we input the same signal to each channel, then no additional information on the joint support is provided from multiple measurements. Therefore, in this worst-case scenario there is no advantage for multiple channels.

In order to capture more closely the true underlying behavior of existing algorithms and observe a performance gain when using several channels, we consider an average analysis. In this setting, the inputs are considered to be random variables so that the case of identical inputs in all channels has zero probability. The idea is to develop conditions on the measurement matrix A such that the inputs can be recovered with high probability given a certain input distribution. Most existing recovery results focus on worst-case analysis. Recently, there have been several papers that consider random ensembles. In [25] random sub-dictionaries of A are considered and analyzed. This allows to obtain results for BP with a single input channel. In [23], average-case performance of single channel thresholding was studied. These ideas were then extended to two multichannel recovery algorithms: thresholding and simultaneous OMP (SOMP) [18, 17]. Under a mild condition on the sparsity and on the matrix A, it was shown that the probability of reconstruction failure decays exponentially with the number of channels. In the present paper we contribute to this line of research by adding an average-case analysis of multichannel BP, that is mixed ℓ_2/ℓ_1-minimization [30, 15, 13, 12].

We denote by A_S the submatrix of A consisting of the columns indexed by $S \subset 1, \ldots, N$, while X^S is the submatrix of X consisting of the rows of X indexed by S. The ℓth column of A is denoted by a_ℓ or A_ℓ. The ℓ_p-norm is denoted by $\|\cdot\|_p$ while $\|\cdot\|_F$ is the Frobenius norm.

2. Multichannel ℓ_1-minimization

We consider multichannel signal recovery where our goal is to recover a jointly-sparse matrix $X \in \mathbb{C}^{N\times L}$ from n linear measurements per channel. Here N denotes the signal length and L the number of channels, i.e., the number of signals. We assume that X is jointly k-sparse, meaning that there are at most k rows in the matrix X that are not identically zero. More formally, we define the support of the matrix X as $\operatorname{supp} X = \bigcup_{\ell=1}^{L} \operatorname{supp} X_\ell$, where the support of the ℓth column is $\operatorname{supp} X_\ell = \{j, X_{j\ell} \neq 0\}$. Our assumption is that $\|X\|_0 = k$ where $\|X\|_0$ is equal to the size of the support. The measurements are given by

$$Y = AX, \quad Y \in \mathbb{C}^{n\times L}, \tag{1}$$

where $A \in \mathbb{C}^{n\times N}$ is a given measurement matrix. Each measurement vector Y_ℓ corresponds to a measurement of the corresponding signal X_ℓ.
The natural approach to determine X given Y is to solve the problem

$$\min_X \|X\|_0 \quad \text{such that} \quad AX = Y. \tag{2}$$

However, (2) is NP hard in general [7]. Therefore, we consider instead the convex relaxation [30, 15, 13] defined by

$$\min \|X\|_{2,1} = \sum_{j=1}^{N} \|X^j\|_2, \quad \text{subject to } AX = Y, \tag{3}$$

which promotes joint sparsity, as argued for instance in [15]. In the single channel case $L = 1$ this is the usual BP principle.

3. Worst Case Recovery Results

Recovery results for the program (3) were considered in [5, 13, 12]. In particular, the lemma below is derived in [5] and follows also from [12].

Proposition 3..1 *Let $S \subset 1, \ldots, N$ and suppose that*

$$\|A_S^\dagger a_\ell\|_1 < 1 \quad \textit{for all } \ell \notin S, \tag{4}$$

with $A_S^\dagger = (A_S^* A_S)^{-1} A_S^*$ denoting the pseudo-inverse of A_S. Then (3) *recovers all $X \in \mathbb{C}^{N\times L}$ with $\operatorname{supp} X = S$ from $Y = AX$.*

Assuming the columns of A are normalized, $\|a_\ell\|_2 = 1$, we can guarantee that (4) holds as long as the coherence μ of A is small enough, where [9]

$$\mu = \max_{j\neq\ell} |\langle a_j, a_\ell\rangle|. \tag{5}$$

The following result follows from Proposition 3..1 or from [12] by noting that the block coherence in this setting is equal to μ/d.

Proposition 3..2 *Assume that*

$$(2k-1)\mu < 1. \tag{6}$$

Then (3) *recovers all X with $\|X\|_0 \leq k$ from $Y = AX$.*

Note that in both of the cited results the conditions do not depend on the number of channels. Indeed, under the same conditions as in Propositions 3..1 and 3..2, it is shown in [26] that BP will recover a single k-sparse vector. Therefore, if (4) holds, then instead of solving (3) we may as well use BP on each of the columns of Y.
The coherence is lower bounded by $\mu \geq \sqrt{\frac{N-n}{n(N-1)}}$ [24]. The lower bound behaves like $1/\sqrt{n}$ for large N, which limits the Proposition 3..2 to maximal sparsities $k = \mathcal{O}(\sqrt{n})$. To improve on this we can generalize existing recovery results [3, 2] based on RIP to the multichannel setup. The next proposition follows from [13]:

Proposition 3..3 *Assume $X \in \mathbb{C}^{n\times N}$ with $\delta_{2k} < \sqrt{2}-1$, where δ_{2k} is the smallest constant δ such that*

$$(1-\delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta)\|x\|_2^2,$$

for all vectors x that are $2k$-sparse. Let $X \in \mathbb{C}^{N\times L}$, $Y = AX$, and let $\overline{X}$ be the minimizer of (3)*. Then*

$$\|X - \overline{X}\|_F \leq Ck^{-1/2}\|X - \hat{X}^{(k)}\|_{1,2}$$

where C is a constant, $\|X\|_F = \sqrt{\operatorname{Tr}(X^*X)}$ is the Frobenius norm of X, $\|X\|_{1,2} = \sum_{j=1}^{N} \|X^j\|_2$, and $\hat{X}^{(k)}$ denotes the best k-term approximation of X, i.e., $\operatorname{supp} \hat{X}^{(k)}$ consists of the indices corresponding to the k largest row norms $\|X^\ell\|_2$. In particular, recovery is exact if $|\operatorname{supp} X| \leq k$.

It is well known that Gaussian and Bernoulli random matrices $A \in \mathbb{R}^{n\times N}$ satisfy $\delta_{2k} \leq \sqrt{2}-1$ with high probability as long as [1, 4]

$$n \geq Ck\log(N/k). \tag{7}$$

Therefore, Proposition 3..3 allows for a smaller number of measurements. However, there is still no dependency on the number of channels. Indeed, under the same RIP condition BP will recover a single k-sparse vector and therefore, as before, BP may as well be applied to each of the columns of Y individually.

4. Average Case Analysis

Intuitively, we would expect multichannel sparse recovery to perform better than single channel recovery. However, in the worst case setting this is not true as already suggested by the results cited above. The reason is very simple. If each channel carries the same signal, $X_\ell = x$ for $\ell = 1, \ldots, L$, then also the components of $Y = AX$ are all the same and we do not have more information on the support of X than provided by a single component Y_ℓ. This can indeed be proven rigorously.

Proposition 4..1 *Suppose there exists a k-sparse vector $x \in \mathbb{R}^N$ that ℓ_1-minimization is not able to recover from $y = Ax$. Then there exists a k-sparse multichannel signal $X \in \mathbb{R}^{N\times L}$ for which mixed ℓ_2/ℓ_1-minimization fails on $Y = AX$.*

For the simple proof we refer to the journal version [14]. Realizing that (3) is not more powerful than usual BP in the worst case, we seek an average-case analysis. This means that we impose a probability model on the k-sparse X. In particular, as in [18], we will assume that on the k-sparse support set S the coefficients of X are independent and follow a normal distribution,

$$X^S = \Sigma\Phi \tag{8}$$

where $\Sigma = \mathrm{diag}(\sigma_j, j \in S) \in \mathbb{R}^{k\times k}$ is an arbitrary diagonal matrix with non-zero diagonal elements σ_j, while $\Phi \in \mathbb{R}^{k\times L}$ is a Gaussian random matrix, i.e., all entries are independent standard normal random variables. Note that taking Σ to be the identity matrix results in a standard Gaussian random matrix, while taking arbitrary non-zero σ_j's on the diagonal of Σ allows for different variances.
The following recovery condition is instrumental in proving average case recovery results for multichannel BP. It generalizes results of [27, 16] for the monochannel case. In order to introduce we need to introduce the sign $\mathrm{sgn}(X)$ of a signal matrix,

$$\mathrm{sgn}(X)_{\ell j} = \begin{cases} \frac{X_{\ell j}}{\|X^\ell\|_2}, & \|X^\ell\|_2 \neq 0; \\ 0, & \|X^\ell\|_2 = 0. \end{cases}$$

Proposition 4..2 *Let $X \in \mathbb{C}^{N\times L}$ with $\mathrm{supp}\, X = S$ and assume A_S to be non-singular. If*

$$\|\,\mathrm{sgn}(X^S)^* A_S^\dagger a_\ell\|_2 < 1 \quad \textit{for all } \ell \notin S \tag{9}$$

then X is the unique minimizer of (3)*.*

Combining the above proposition with a concentration inequality for sums of independent random variables that are uniformly distributed on the sphere [19], we arrive at the following average case recovery result for multichannel BP.

Theorem 4..3 *Let $S \subset \{1,\dots,N\}$ be a set of cardinality k and let $X \in \mathbb{R}^{N\times L}$ with $\mathrm{supp}\, X \subset \{1,\dots,N\}$ such that the coefficients on S are given by* (8) *with some diagonal matrix $\Sigma \in \mathbb{R}^{k\times k}$. If*

$$\|A_S^\dagger a_\ell\|_2 \leq \alpha < 1 \quad \textit{for all } \ell \notin S, \tag{10}$$

then with probability at least

$$1 - N\exp\left(-\frac{L}{2}(\alpha^{-2} - \log(\alpha^{-2}) - 1)\right) \tag{11}$$

(3) *recovers X from $Y = AX$.*

The proof of the theorem will appear in the journal version [14]. For $\alpha < 1$ we are guaranteed that the exponent has a negative argument, and therefore the error decays exponentially in L. We note that for the monochannel case $L = 1$, Theorem 4..3 is contained implicitly in [28, Theorem 13]. The appearance of the 2-norm in (10) instead of the 1-norm as in (4) makes the condition of the theorem weaker than worst-case estimates.
Let us finally state conditions on the matrix A and the sparsity level k ensuring that $\|A_S^\dagger a_\ell\|_2$ is small, which is needed in order to apply Theorem 4..3.

Proposition 4..4 *Suppose A has restricted isometry constant $\delta_{k+1} \leq \delta < 1/2$. If $S \subset \{1,\dots,N\}$ has cardinality k then*

$$\|A_S^\dagger a_\ell\|_2 \leq \frac{\delta}{1-\delta} < 1 \quad \textit{for all } \ell \notin S.$$

Note that in contrast to the worst case result in Proposition 3..3 where a condition on δ_{2k} is needed, we only require that δ_{k+1} is small, which is clearly weaker. For random matrices A we have the following bound on $\|A_S^\dagger a_\ell\|_2$.

Proposition 4..5 *Let $S \subset \{1,\dots,N\}$ be a set of cardinality k and suppose that $A \in \mathbb{R}^{n\times N}$ is drawn at random according to a Gaussian or Bernoulli distribution. Then*

$$\|A_S^\dagger a_\ell\|_2 \leq \delta \quad \textit{for all } \ell \notin S$$

with probability at least $1-\epsilon$ provided that

$$n \geq C\delta^{-2}[(k+1)\ln(1+12/\delta) + \ln(2N/\epsilon)]. \tag{12}$$

The constant C is no larger than $162/7 \approx 23.1$.

Note that the log-factor in (12) enters only as an additive term, while in (7) it appears as multiplicative factor.

5. Conclusion

Our main result is that under mild conditions on the sparsity and measurement matrix, the probability of failure of multichannel BP (3) decays exponentially with the number of channels. To develop this result we assumed a probability model on the non-zero coefficients of a jointly sparse signal. This shows that multichannel BP outperforms single channel BP applied to each channel individually, on average. Proofs of our theorems, together with improved results for simple thresholding and numerical experiments will appear in [14].

6. Acknowledgements

The work of YE was supported in part by the Israel Science Foundation under Grant no. 1081/07 and by the European Commission in the framework of the FP7 Network of Excellence in Wireless COMmunications NEWCOM++ (contract no. 216715). HR acknowledges funding by the Hausdorff Center for Mathematics, University of Bonn and the WWTF project SPORTS (MA 07-004).

References:

[1] R. G. Baraniuk, M. Davenport, R. A. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constr. Approx.*, 28(3):253–263, 2008.

[2] E. J. Candès. The restricted isometry property and its implications for compressed sensing. *Compte Rendus de l'Academie des Sciences, Paris, Serie I*, 346:589–592, 2008.

[3] E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1207–1223, 2006.

[4] E. J. Candès and T. Tao. Near optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theory*, 52(12):5406–5425, 2006.

[5] J. Chen and X. Huo. Theoretical results on sparse representations of multiple-measurement vectors. *IEEE Trans. Signal Processing*, 54(12):4634–4643, Dec. 2006.

[6] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado. Sparse solutions to linear inverse problems with multiple measurement vectors. *IEEE Trans. Signal Processing*, 53(7):2477–2488, July 2005.

[7] G. Davis, S. Mallat, and M. Avellaneda. Adaptive greedy approximations. *Constr. Approx.*, 13(1):57–98, 1997.

[8] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[9] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions Info. Theory*, 47(7):2845–2862, 2001.

[10] David L. Donoho. For most large underdetermined systems of linear equations the minimal l^1 solution is also the sparsest solution. *Commun. Pure Appl. Anal.*, 59(6):797–829, 2006.

[11] Y. C. Eldar. Compressed sensing of analog signals. submitted to *IEEE Trans. Signal Processing*.

[12] Y. C. Eldar and H. Bölcskei. Block-sparsity: Coherence and efficient recovery. to appear in *ICASSP09*.

[13] Y. C. Eldar and M. Mishali. Robust recovery of signals from a union of subspaces. submitted to *IEEE Trans. Inf. Theory*.

[14] Y. C. Eldar and H. Rauhut. Average case analysis for multichannel sparse recovery using convex relaxation. *preprint*, 2009.

[15] M. Fornasier and H. Rauhut. Recovery algorithms for vector valued data with joint sparsity constraints. *SIAM J. Numer. Anal.*, 46(2):577–613, 2008.

[16] J. J. Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Trans. Inform. Theory*, 50(6):1341–1344, 2004.

[17] R. Gribonval, B. Mailhe, H. Rauhut, K. Schnass, and P. Vandergheynst. Average case analysis of multichannel thresholding. In *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 2007.

[18] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst. Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms. *J. Fourier Anal. Appl.*, 14(5):655–687, 2008.

[19] H. König and S. Kwapień. Best Khintchine type inequalities for sums of independent, rotationally invariant random vectors. *Positivity*, 5(2):115–152, 2001.

[20] M. Mishali and Y. C. Eldar. Reduce and boost: Recovering arbitrary sets of jointly sparse vectors. *IEEE Trans. Signal Process.*, 56(10):4692–4702, Oct. 2008.

[21] M. Mishali and Y. C. Eldar. Blind multi-band signal reconstruction: Compressed sensing for analog signals. *IEEE Trans. Signal Process.*, 57(3):993–1009, Mar. 2009.

[22] H. Rauhut. On the impossibility of uniform sparse reconstruction using greedy methods. *Sampl. Theory Signal Image Process.*, 7(2):197–215, 2008.

[23] K. Schnass and P. Vandergheynst. Average performance analysis for thresholding. *IEEE Signal Processing Letters*, 14(11):828–831, Nov. 2007.

[24] T. Strohmer and R. W. Heath. Grassmannian frames with applications to coding and communication. *Appl. Comput. Harmon. Anal.*, 14(3):257–275, 2003.

[25] G. Teschke. Multi-frame representations in linear inverse problems with mixed multi-constraints. *Appl. Comput. Harmon. Anal.*, 22(1):43–60, 2007.

[26] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10):2231–2242, 2004.

[27] J. A. Tropp. Recovery of short, complex linear combinations via l_1 minimization. *IEEE Trans. Inform. Theory*, 51(4):1568–1570, 2005.

[28] J. A. Tropp. On the conditioning of random subdictionaries. *Appl. Comput. Harmon. Anal.*, to appear.

[29] J. A. Tropp, A. C. Gilbert, and M. J. Strauss. Algorithms for simultaneous sparse approximation: part I: Greedy pursuit. *Signal Processing*, 86(3):572 – 588, 2006.

[30] J. A. Tropp, A. C. Gilbert, and M. J. Strauss. Algorithms for simultaneous sparse approximation: part II: Convex relaxation. *Signal Processing*, 86(3):589 – 602, 2006.

Special session on

Sampling
Using
Finite Rate of Innovation Principles

Chairs: Pier-Luigi DRAGOTTI, Pina MARZILIANO

Sampling of Sparse Signals in Fractional Fourier Domain

Ayush Bhandari [1] and Pina Marziliano [2]

(1)Temasek Labs @ NTU, 50 Nanyang Drive, Singapore - 637553
(2) School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore - 639798
{ayushbhandari,epina}@ntu.edu.sg

Abstract: In this paper, we formulate the problem of sampling sparse signals in fractional Fourier domain. The fractional Fourier transform (FrFT) can be seen as a generalization of the classical Fourier transform. Extension of Shannon's sampling theorem to the class of signals which are fractional bandlimited shows its association to a Nyquist-like bound. Thus proving that signals that have a non-bandlimited representation in FrFT domain cannot be sampled. We prove that under suitable conditions, it is possible to sample sparse (in time) signals by using the *Finite Rate of Innovation* (FRI) signal model. In particular, we propose a uniform sampling and reconstruction procedure for a periodic stream of Diracs, which have a non-bandlimited representation in FrFT domain. This generalizes the FRI sampling and reconstruction scheme in the Fourier domain to the FrFT domain.

1. Introduction

Shannon's sampling theorem [1] provides access to the digital world. Our understanding of this sampling theorem together with the reconstruction formula is solely based on the frequency content of the signal of interest. This is where the indispensable Fourier transform comes into the picture.

Almeida [2] introduced the fractional Fourier transform or the FrFT–a generalization of the Fourier transform–to the signal processing community in 1994. The generalization of the Fourier transform by FrFT has several interesting consequences from the signal processing perspective. For instance, non-bandlimited signals in the Fourier domain can still have a compactly supported representation in FrFT domain [3], when dealing with non stationary distortions, the FrFT based filters can perform better than Fourier domain based filters (in sense of mean square error) [4] etc. To give the reader an idea about the growing popularity of FrFT, it would be worth mentioning that on at least eight occasions including, [3, 5, 6, 7, 8, 9, 10, 11], Shannon's sampling theorem [1, 12] was independently extended to the class of fractional bandlimited signals. In [13], the FrFT of a signal or a function, say $x(t)$, is defined by

$$\widehat{x}_\theta(\omega) = \text{FrFT}\{x(t)\} = \int x(t) K_\theta(t,\omega) dt \qquad (1)$$

where

$$K_\theta(t,\omega) \overset{\text{def}}{=} \begin{cases} \sqrt{\frac{1-j\cot\theta}{2\pi}} e^{j\frac{t^2+\omega^2}{2}\cot\theta - j\omega t\csc\theta}, & \theta \neq p\pi \\ \delta(t-\omega), & \theta = 2p\pi \\ \delta(t+\omega), & \theta + \pi = 2p\pi \end{cases} \qquad (2)$$

is the transformation kernel, parametrized by the fractional order $\theta \in \mathbb{R}$ and p is some integer.

The FrFT of a time-frequency representation e.g. Gabor Transform results in rotation of the plane by the fractional order of the FrFT [2]. Thus, we denote fractional order by θ and from now on, we will use fractional order and angle interchangeably. The inverse-FrFT with respect to angle θ is the FrFT at angle $-\theta$, given by,

$$x(t) = \int_{-\infty}^{\infty} \widehat{x}_\theta(\omega) K_{-\theta}(t,\omega) d\omega. \qquad (3)$$

Whenever $\theta = \pi/2$, (1) collapses to the classical Fourier transform definition. A direct consequence of the generalization of the Fourier transform by the FrFT results in a modification in the idea of bandlimitedness. Its impact is visible in the change that manifests in Shannon's sampling theorem for fractional bandlimited signals [11], which is stated in Theorem 1.

Theorem 1 (Shannon–FrFT)**.** *Let $x(t)$ be a continuous-time signal. If the spectrum of $x(t)$, i.e. $\widehat{x}_\theta(\omega)$ is fractional bandlimited to ω_m which means, $\widehat{x}_\theta(\omega) = 0$, when $|\omega| > \omega_m$, then $x(t)$ is completely determined by giving its ordinates at a series of equidistant points spaced $T = \frac{\pi}{\omega_m}\sin\theta$ seconds apart.*

This theorem has an equivalence to the Shannon's sampling theorem for $\theta = \pi/2$. The reconstruction formula for fractional bandlimited signals is given in [11],

$$x(t) = \lambda_\theta^*(t) \sum_{n\in\mathbb{Z}} \lambda_\theta(nT)\, x(nT) \text{sinc}\left((t-nT)\,\omega_m \csc\theta\right) \qquad (4)$$

where $\lambda_\theta(\cdot) \overset{\text{def}}{=} e^{j(\cdot)^2\frac{\cot\theta}{2}}$ is a domain independent chirp modulation function and the '*' in the superscript denotes complex conjugation. If $\widetilde{x}(t)$ is the approximation of $x(t)$, then $\|\widetilde{x}(t) - x(t)\|^2 = 0$ when $\omega_m \leqslant \frac{\omega_s}{2}\sin\theta$–the Nyquist rate for FrFT–where $\omega_s = 2\pi/T$ is the sampling frequency. Note that all the aforementioned results are equivalent to Shannon's sampling theorem with respect to

Fourier domain for $\theta = \pi/2$. Theorem 1 (for FrFT) has a striking similarity with the Shannon's sampling theorem (for FT), in that, sampling non-bandlimited signals is impossible. Consider Dirac's delta function or $\delta(t)$. Using (2), we have,

$$\widehat{\delta}_\theta(\omega) = \text{FrFT}\left\{\delta\left(t\right)\right\} = \sqrt{\tfrac{1-j\cot\theta}{2\pi}}\lambda_\theta\left(\omega\right) \qquad (5)$$

which is a non-bandlimited function (and least sparse when compared to the time-domain counterpart) and thus, Theorem 1 fails to answer the following question: If $x(t)$ is a fractional non-bandlimited signal, then, how can we sample and reconstruct such a signal? To make this statement clear, we introduce the fractional convolution operator, which is denoted by '$*_\theta$'. Accordingly, filtering $x(t)$ by a filter, $h(t)$, in 'fractional sense' [1] is equivalent to [14],

$$x(t)*_\theta h(t) = \sqrt{\tfrac{1-j\cot\theta}{2\pi}}\lambda_\theta^*(t)\cdot([x(t)\lambda_\theta(t)] * [h(t)\lambda_\theta(t)]) \qquad (6)$$

where '$*$' denotes the usual convolution operator. In light of this definition, we wish to address the problem of recovering *parsimonious* $x(t)$ from the samples of its filtered version, i.e., $y(nT) = x(t) *_\theta h(t)|_{t=nT}, n \in \mathbb{Z}$. This problem has a natural/strong link with that of sparse sampling [15, 16, 17]. The Heisenberg-Gabor uncertainty principle for the FrFT [18] (a generalization of the Fourier duality) asserts that the product of spreads of $\widehat{x}_\theta\left(\omega\right)$ and $x(t)$ has a lower bound which is proportional to $\frac{\sin^2\theta}{4}$ (assuming that $\|x\| = 1$). This implies that sparsity in one domain will lead to loss of compact support in canonically conjugate domain.

Our contribution in this article is to propose a sampling and reconstruction scheme for signals which have a sparse representation in time domain and whose fractional spectrum is non-bandlimited. We model our sparse signal as a continuous periodic stream of Diracs which is being observed by an acquisition device which deploys a sinc-based filter.

The paper is organized as follows: We assume that the reader is familiar with basic ideas outlined in [12, 16, 17]. In Section II, we introduce our sparse signal model and the definition of the fractional Fourier series (FrFS). Using these as preliminaries, in Section III, we derive an equivalent representation of our signal in FrFT domain. In Section IV, we discuss the sampling theorem and its completeness and Section V is the conclusion.

2. Preliminaries

2.1 Sparse Signal Model

We model our sparse signal as a periodic stream of K Diracs, i.e.

$$x(t) = \sum_{k=0}^{K-1} c_k \sum_{n\in\mathbb{Z}} \delta(t - t_k - n\tau) \qquad (7)$$

with period τ, weights $\{c_k\}_{k=0}^{K-1}$ and arbitrary shifts, $\{t_k\}_{k=0}^{K-1} \subset [0,\tau)$. In sense of [16], the signal has $2K$ degrees of freedom per period and the rate of innovation being $\rho = \frac{2K}{\tau}$. From now on, the signal $x(t)$ will denote the stream of Diracs.

2.2 Fractional Fourier Series (FrFS)

Periodic signals can be expanded in FrFT domain as a fractional Fourier series or FrFS [19]. The FrFS of a periodic signal, say $x(t)$, can be written as,

$$x(t) = \sum_{m\in\mathbb{Z}} \widehat{x}_\theta[m]\Phi_\theta(m,t) \qquad (8)$$

where,

$$\Phi_\theta^*(m,t) = \sqrt{\frac{\sin\theta - j\cos\theta}{\tau}} e^{j\frac{t^2+(2\pi m\sin\theta/\tau)^2}{2}\cot\theta - j2\pi mt/\tau}$$

constitutes the basis for FrFS expansion for a τ-periodic $x(t)$. The FrFS coefficients are given by,

$$\widehat{x}_\theta[m] = \int_{\langle\tau\rangle} x(t)\Phi_\theta^*(m,t)dt = \langle x, \Phi_\theta(m,\cdot)\rangle \qquad (9)$$

where $\langle\tau\rangle$ denotes the integral width and $\langle a,b\rangle = \int a(t)b^*(t)dt$ denotes the inner product. The well-known Fourier series (FS) is just a special case of FrFS for $\theta = \frac{\pi}{2}$.

3. Stream of Diracs in Fractional Fourier Domain

In Fourier analysis, the Poisson summation formula (PSF) plays an important role. It is a well-known fact that a stream of Diracs (Dirac comb) in time-domain is another stream of Diracs in Fourier domain. In this subsection, we will derive the equivalent representation of Dirac comb in FrFT domain. This can be seen as a generalization of the Poisson summation formula for Dirac comb in FrFT domain.

Theorem 2. *Let $\sum_{n\in\mathbb{Z}}\delta(t-n\tau)$ be a Dirac comb, then*

$$\sum_{n\in\mathbb{Z}} \delta\left(t - n\tau\right) \overset{FrFT}{\longleftrightarrow}$$

$$\frac{1}{\tau}\sqrt{\tfrac{2\pi}{1-j\cot\theta}} \sum_{k\in\mathbb{Z}} \widehat{\delta}_\theta\left[k\omega_0\sin\theta\right] e^{-j\left(t^2+\frac{(k\omega_0\sin\theta)^2}{2}\right)\cot\theta + jk\omega_0 t}$$

where $\omega_0 = \frac{2\pi}{\tau}$.

Proof. Let $s(t) \overset{\text{def}}{=} \sum_{n\in\mathbb{Z}}\delta(t-n\tau)$. The proof is done by expanding $s(t)$ in FrFS basis or,

$$s(t) = \sum_{k\in\mathbb{Z}} \underbrace{\langle s, \Phi_\theta\rangle}_{\widehat{s}_\theta[k]} \Phi_\theta\left(k,t\right). \qquad (10)$$

[1] We adhere to this modified definition of convolution operator as it inherits the fractional Fourier duality property, in that, FrFT $\{x(t) *_\theta h(t)\} = \lambda_\theta^*(\omega)\cdot\widehat{x}_\theta(\omega)\widehat{h}_\theta(\omega)$, which does not hold for the FrFT of $x(t) * h(t)$ unless $\theta = \frac{\pi}{2}$.

The coefficients of this expansion are given by,

$$\begin{aligned}
\widehat{s}_\theta[k] &\overset{(9)}{=} \langle s, \Phi_\theta(k,t)\rangle \\
&= \frac{\kappa(\theta)}{\sqrt{\tau}} \int_{t_0}^{t_0+\tau} s(t)\,\Phi_\theta^*(k,t)\,dt, \quad \forall t_0 \in \mathbb{R} \\
&= \frac{\kappa(\theta)}{\sqrt{\tau}} \int_{-\tau/2}^{\tau/2} \delta(t)\, e^{j\left(t^2+(k\omega_0\sin\theta)^2/2\right)\cot\theta - jk\omega_0 t} dt \\
&\quad \left(\text{since } s(t+\tau)=s(t) \text{ and } s(t)=\delta(t), t\in\left[\tfrac{-\tau}{2},\tfrac{\tau}{2}\right]\right) \\
&= \frac{\kappa(\theta)}{\sqrt{\tau}} e^{j\left((k\omega_0\sin\theta)^2/2\right)\cot\theta} \\
&\overset{(5)}{=} \frac{\kappa(\theta)}{\sqrt{\tau}} \sqrt{\frac{2\pi}{1-j\cot\theta}}\,\widehat{\delta}_\theta[k\omega_0\sin\theta] \qquad (11)
\end{aligned}$$

where $\kappa(\theta)=\sqrt{\sin\theta - j\cos\theta}$. Back substitution of (11) in (10) results in,

$$\begin{aligned}
s(t) &= \frac{1}{\tau}\sqrt{\tfrac{2\pi}{1-j\cot\theta}} \\
&\times \sum_{k\in\mathbb{Z}} \widehat{\delta}_\theta[k\omega_0\sin\theta]\, e^{-j\left(t^2+\frac{(k\omega_0\sin\theta)^2}{2}\right)\cot\theta + jk\omega_0 t}.
\end{aligned}$$

This concludes the proof. ■

For sake of convenience, we will assume that the constant $\sqrt{\frac{1-j\cot\theta}{2\pi}}$ has been absorbed in τ. Note that at $\theta=\frac{\pi}{2}$, $s(t)=\frac{1}{\tau}\sum_{k\in\mathbb{Z}} e^{jk\omega_0 t}$ which is the result of applying the PSF on $s(t)$ in Fourier domain. Our immediate goal now is to derive the FrFS equivalent of $x(t)$ in (7). Since $x(t)$ is a linear combination of some $s(t)$ delayed by some time shift t_k, it will be useful to recall shift property of FrFT [2] which states that,

$$\begin{aligned}
&\text{FrFT}\{s(t-t_k)\} \\
&= \widehat{s}_\theta(\omega - t_k\cos\theta)\, e^{j\frac{1}{2}t_k^2\sin\theta\cos\theta - j\omega t_k\sin\theta}. \qquad (12)
\end{aligned}$$

Therefore, call $x(t)=\sum_{k=0}^{K-1} c_k\cdot s_k(t)$ where $s_k(t)$ is the time-shifted version of $s(t)$ with shift parameter t_k. Using Theorem 2 and the shift-property of FrFT, we have,

$$\begin{aligned}
s_k(t) &= \sum_{n\in\mathbb{Z}} \delta(t-t_k-n\tau) \\
&\overset{(8)}{=} \sum_{m\in\mathbb{Z}} \text{FrFT}\{\,\delta(t-t_k)\}\,|_{\omega=m\omega_0\sin\theta}\,\Phi_\theta(m,t) \\
&\overset{(12)}{=} \underbrace{\frac{1}{\tau}\sum\nolimits_{m\in\mathbb{Z}} e^{j\frac{\cot\theta}{2}\left(t_k^2-t^2\right)+jm\omega_0(t-t_k)}}_{\text{PSF for Dirac Comb in FrFT}}.
\end{aligned}$$

Having obtained the FrFT-version of $s_k(t)$, we can write,

$$\begin{aligned}
x(t) &= \sum_{k=0}^{K-1} c_k \cdot \sum_{n\in\mathbb{Z}} \delta(t-t_k-n\tau) \\
&= \sum_{k=0}^{K-1} c_k \sum_{m\in\mathbb{Z}} e^{j\frac{\cot\theta}{2}\left(t_k^2-t^2\right)+jm\omega_0(t-t_k)} \\
&= e^{-jt^2\frac{\cot\theta}{2}} \sum_{m\in\mathbb{Z}} \frac{1}{\tau} \underbrace{\left(\sum_{k=0}^{K-1} c_k e^{j\frac{\cot\theta}{2}\left(t_k^2\right)-jm\omega_0 t_k}\right)}_{p[m]} e^{j\frac{2\pi m}{\tau}t}.
\end{aligned}$$

Note that $x(t)$ is non-bandlimited, however, it can be completely described by the knowledge of $p[m]$ which in turn can be expanded as a linear combination of K complex exponentials.

4. Sampling and Reconstruction of Sparse Signals in Fractional Fourier Domain

We assume that a sinc–based kernel is used to prefilter $x(t)$. In particular, we let the sampling kernel to be $\varphi_n(t) = e^{-j\frac{\cot\theta}{2}t^2}\,\text{sinc}(t-nT)$. Integer translates of $\varphi_n(t)$ form an orthonormal basis and the FrFT of $\varphi(t)(=\varphi_0(t))$ is given by $\widehat{\varphi}_\theta(\omega) = \sqrt{\frac{1-j\cot\theta}{2\pi}}\left(e^{-j\frac{\cot\theta}{2}\omega^2}\right)\text{rect}(\omega/2\pi)$. In light of the definition in (6), prefiltering the input signal $x(t)$ with the kernel/low-pass filter $\varphi(-t)$ and sampling can be written as, $y(nT) = x(t)*_\theta \varphi(-t)|_{t=nT}$. The main result is in the form of the following theorem.

Theorem 3. *Let $x(t)$ be a τ-periodic stream of Diracs weighted by coefficients $\{c_k\}_{k=0}^{K-1}$ and locations $\{t_k\}_{k=0}^{K-1}$ with finite rate of innovation $\rho=\frac{2K}{\tau}$. Let the sampling kernel/prefilter $\varphi(t)$ be an ideal low-pass filter which has fractional bandwidth $[-B\pi, B\pi]$, where B is chosen such that $B\geq\rho$. If the filtered version of $x(t)$, i.e. $y(t)=x(t)*_\theta\varphi(-t)$ is sampled uniformly at locations $t=nT$, $n=0,\ldots,N-1$ then the samples,*

$$y(nT) = x(t)*_\theta\varphi(-t)|_{t=nT}, n=0,\ldots,N-1,$$

are a sufficient characterization of $x(t)$, provided that $N\geq 2M_\theta+1$ and $M_\theta=\left\lfloor\frac{B\tau\csc\theta}{2}\right\rfloor$.

Proof. Using the following FrFT pair,

$$\begin{aligned}
&\sqrt{\tfrac{1-j\cot\theta}{2\pi}}\lambda_\theta^*(\omega)\cdot\text{rect}(\tfrac{\omega}{2\pi B}) \overset{\text{FrFT}}{\longleftrightarrow} \\
&(B\csc\theta)\,\lambda_\theta^*(t)\,\text{sinc}(Bt\csc\theta)
\end{aligned}$$

we define our sampling kernel as,

$$\varphi_B(t-nT) = \lambda_\theta^*(t)\,\varphi(B\csc\theta\,(t-nT))$$

which is compactly supported over $[-B\pi, B\pi]$. Prefiltering and sampling $x(t)$ results in,

$$\begin{aligned}
y(nT) &= x(t)*_\theta\varphi(-t)|_{t=nT}, n=0,\ldots,N-1 \\
&= \frac{\lambda_\theta^*(nT)}{\tau}\sum_{m\in\mathbb{Z}} p[m] \\
&\times\left\langle e^{j\frac{2\pi m}{\tau}t}, (B\csc\theta)\,\text{sinc}((B\csc\theta)(t-nT))\right\rangle.
\end{aligned}$$

The inner product in the above step is further simplified using the Fourier integral,

$$\begin{aligned}
&\left\langle e^{j\frac{2\pi m}{\tau}t}, (B\csc\theta)\,\text{sinc}((B\csc\theta)(t-nT))\right\rangle = \\
&\qquad \text{rect}(\tfrac{m}{B\tau\csc\theta}) e^{j\frac{2\pi m}{\tau}(nT)}.
\end{aligned}$$

We can therefore conclude that,

$$\begin{aligned}
y(nT) &= \frac{\lambda_\theta^*(nT)}{\tau}\sum_{m\in\mathbb{Z}} p[m]\,\text{rect}(\tfrac{m}{B\tau\csc\theta}) e^{j\frac{2\pi m}{\tau}(nT)} \\
&= \frac{\lambda_\theta^*(nT)}{\tau}\sum_{m=-M_\theta}^{M_\theta} p[m] e^{j\frac{2\pi m}{\tau}(nT)}, n=0,\ldots,N-1
\end{aligned}$$

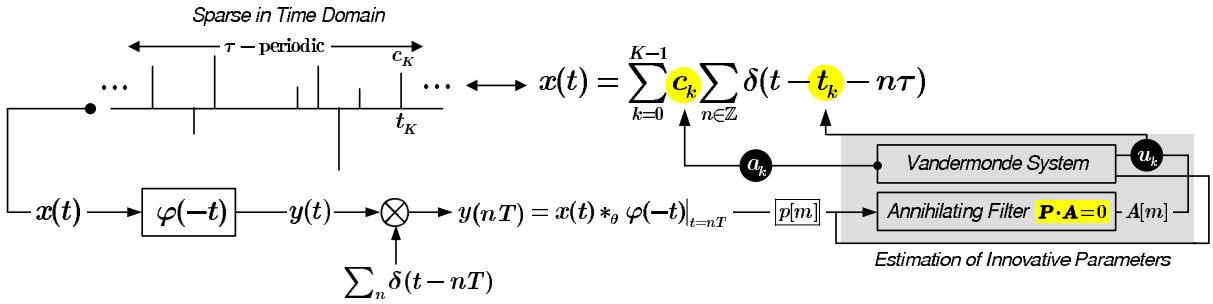


Figure 1: Sampling and reconstruction of periodic stream of Diracs in FrFT domain.

where $M_\theta = \lfloor \frac{B\tau \csc\theta}{2} \rfloor$.

Signal reconstruction from its samples: Call $p[m] = \sum_{k=0}^{K-1} a_k u_k^m$ – a linear combination of K-complex exponentials, $u_k = \lambda^*_{\pi/2}\left(\sqrt{\omega_0 t_k}\right)$ with weights $a_k = c_k \cdot \lambda_\theta(t_k)$. The problem of calculating $\{a_k\}_{k=0}^{K-1}$ and $\{u_k\}_{k=0}^{K-1}$ is based on finding a suitable polynomial $A(z) = \prod_{k=0}^{K-1}\left(1 - u_k z^{-1}\right)$ whose inverse z-transform yields the annihilating filter coefficients, $A[m]$ which annihilate $p[m]$. In matrix notation, finding $A[m]$ is equivalent to finding a corresponding vector $\mathbf{A}$ that forms a null space of a suitable submatrix of $p[m]$ i.e. $\mathbf{P}^{(2M_\theta - K+1)\times(K+1)}$ – which is essentially the set $\text{Null}(\mathbf{P}) = \left\{\mathbf{A} \in \mathbb{R}^{K+1} : \mathbf{P}\cdot\mathbf{A} = \mathbf{0}\right\}$. For details of this computation, the reader is referred to (cf. Pg. 1427, [16]). Figure 1 shows the layout of this algorithm. ■

5. Conclusion

We presented a scheme for sampling and reconstruction of sparse signals in fractional Fourier domain. A direct consequence of modeling our signal of interest as a *Finite Rate of Innovation* signal, is that, the outcome bears an acute resemblance with the results previously derived, for the Fourier domain case. This simplifies the problem to the extent that reconstruction strategy remains unchanged and as we have shown, one can obtain the precise locations and amplitudes of the stream of Diracs using the annihilating filter method. Since time and frequency domains are special cases of the FrFT domain, it turns out that the number of values (M_θ) required for exact reconstruction of time domain signal depends on the chirp rate of transformation, i.e. θ.

References:

[1] C. E. Shannon. Communications in the presence of noise. *Proc. of the IRE*, 37:10–21, January 1949.

[2] L. B. Almeida. The fractional Fourier transform and time-frequency representations. *IEEE Trans. Signal Proc.*, 42(11):3084–3091, Nov 1994.

[3] X. G. Xia. On bandlimited signals with fractional Fourier transform. *IEEE Signal Proc. Letters*, 3(3):72–74, Mar 1996.

[4] A. Kutay, H. M. Ozaktas, O. Ankan, and L. Onural. Optimal filtering in fractional Fourier domains. *IEEE Trans. Signal Proc.*, 45(5):1129–1143, May 1997.

[5] A. I. Zayed. On the relationship between the Fourier and fractional Fourier transforms. *IEEE Signal Proc. Letters*, 3(12):310–311, Dec 1996.

[6] T. Erseghe, P. Kraniauskas, and G. Carioraro. Unified fractional Fourier transform and sampling theorem. *IEEE Trans. Signal Proc.*, 47(12):3419–3423, Dec 1999.

[7] A. I. Zayed and A. G. García. New sampling formulae for the fractional Fourier transform. *Signal Proc.*, 77(1):111–114, 1999.

[8] A. G. García. Orthogonal sampling formulas: A unified approach. *SIAM Rev.*, 42(3):499–512, 2000.

[9] Ç. Candan and H. M. Ozaktas. Sampling and series expansion theorems for fractional Fourier and other transforms. *Signal Proc.*, 83(11):2455–2457, 2003.

[10] R. Torres, P. F. Pellat, and Y. Torres. Sampling theorem for fractional bandlimited signals: A self-contained proof. application to digital holography. *IEEE Signal Proc. Letters*, 13(11):676–679, Nov. 2006.

[11] R. Tao, B. Deng, Z.-Q. Wei, and Y. Wang. Sampling and sampling rate conversion of band limited signals in the fractional Fourier transform domain. *IEEE Trans. Signal Proc.*, 56(1):158–171, Jan. 2008.

[12] M. Unser. Sampling-50 years after Shannon. *Proc. IEEE*, 88(4):569–587, 2000.

[13] H. M. Ozaktas and M. A. Kutay. *Introduction to the fractional Fourier transform and its applications*. Academic Press, 1999.

[14] P. Kraniauskas, G. Cariolaro, and T. Erseghe. Method for defining a class of fractional operations. *IEEE Trans. Signal Proc.*, 46(10):2804–2807, Oct 1998.

[15] P. Marziliano. *Sampling innovations*. PhD thesis, EPFL, Switzerland, 2001.

[16] M. Vetterli, P. Marziliano, and T. Blu. Sampling signals with finite rate of innovation. *IEEE Trans. Signal Proc.*, 50(6):1417–1428, Jun 2002.

[17] T. Blu, P.-L. Dragotti, M. Vetterli, P. Marziliano, and L. Coulot. Sparse sampling of signal innovations. *IEEE Signal Proc. Mag.*, 25(2):31–40, March 2008.

[18] S. Shinde and V. M. Gadre. An uncertainty principle for real signals in the fractional Fourier transform domain. *IEEE Trans. Signal Proc.*, 49(11):2545–2548, Nov 2001.

[19] S. C. Pei, M. H. Yeh, and T. L. Luo. Fractional Fourier series expansion for finite signals and dual extension to discrete-time fractional Fourier transform. *IEEE Trans. Signal Proc.*, 47(10):2883–2888, Oct 1999.

Estimating Signals With Finite Rate of Innovation From Noisy Samples: A Stochastic Algorithm

Vincent Y. F. Tan and Vivek K Goyal

Massachusetts Institute of Technology, Cambridge, MA 02139 USA
vtan@mit.edu, vgoyal@mit.edu

Abstract:

As an example of the concept of rate of innovation, signals that are linear combinations of a finite number of Diracs per unit time can be acquired by linear filtering followed by uniform sampling. However, in reality, samples are not noiseless. In a recent paper, we introduced a novel *stochastic* algorithm to reconstruct a signal with finite rate of innovation from its *noisy* samples. Even though variants of this problem has been approached previously, satisfactory solutions are only available for certain classes of sampling kernels, for example kernels which satisfy the Strang–Fix condition. In our paper, we considered the infinite-support Gaussian kernel, which does not satisfy the Strang–Fix condition. Other classes of kernels can be employed. Our algorithm is based on Gibbs sampling, a Markov chain Monte Carlo (MCMC) method. This paper summarizes the algorithm and provides numerical simulations that demonstrate the accuracy and robustness of our algorithm.

1. Introduction

The celebrated Nyquist–Shannon sampling theorem [4, 6] states that a signal $x(t)$ known to be bandlimited to $\Omega_{\max}$ Hz is uniquely determined by samples of $x(t)$ spaced $1/(2\Omega_{\max})$ sec apart. The textbook reconstruction procedure is to feed the samples as impulses to an ideal lowpass (sinc) filter. Furthermore, if $x(t)$ is not bandlimited or the samples are noisy, introducing pre-filtering by the appropriate sinc *sampling kernel* gives a procedure that finds the orthogonal projection to the space of $\Omega_{\max}$-bandlimited signals. Thus the noisy case is handled by simple, linear, time-invariant processing.

Sampling has come a long way since the sampling theorem, but until recently the results have mostly applied only to signals contained in shift-invariant subspaces [9]. Moving out of this restrictive setting, Vetterli *et al.* [10] showed that it is possible to develop sampling schemes for certain classes of non-bandlimited signals that are not subspaces. As described in [10], for reconstruction from samples it is necessary for the class of signals to have *finite rate of innovation* (FRI). The paradigmatic example is the class of signals expressed as

$$x(t) = \sum_k c_k \phi(t - t_k) \tag{1}$$

where $\phi(t)$ is some known function. For each term in the sum, the signal has two real parameters c_k and t_k. If the density of t_ks (the number that appear per unit of time) is finite, the signal has FRI. It is shown constructively in [10] that the signal can be recovered from (noiseless) uniform samples of $x(t)*h(t)$ (at a sufficient rate) when $\phi(t)*h(t)$ is a sinc or Gaussian function. Results in [2] are based on similar reconstruction algorithms and greatly reduce the restrictions on the sampling kernel $h(t)$.

In practice, though, acquisition of samples is not a noiseless process. For instance, an analog-to-digital converter (ADC) has several sources of noise, including thermal noise, aperture uncertainty, comparator ambiguity, and quantization [11]. Hence, samples are inherently noisy. This motivates our central question: *Given the signal model (i.e. a signal with FRI) and the noise model, how well can we approximate the parameters that describe the signal and hence the signal itself?* In this work, we address this question by developing a novel algorithm to reconstruct the signal from the noisy samples. The main contribution is to show that a stochastic approach can effectively circumvent the ill-conditioning of algebraic techniques.

This paper is an abridged version of [7], where many additional details can be found.

2. Problem Definition and Notation

The basic setup is shown in Fig. 1. As mentioned in the introduction, we consider a class of signals characterized by a finite number of parameters. In this paper, similar to [2, 3, 10], the class is the weighted sum of K Diracs

$$x(t) = \sum_{k=1}^{K} c_k \delta(t - t_k). \tag{2}$$

(The use of a Dirac delta simplifies the discussion. It can be replaced by a known pulse $\phi(t)$ and then absorbed into the sampling kernel $h(t)$, yielding an effective sampling kernel $\phi(t) * h(t)$.) The signal to be estimated $x(t)$ is filtered using a Gaussian lowpass filter

$$h(t) = \exp\left(-\frac{t^2}{2\sigma_h^2}\right) \tag{3}$$

with width σ_h to give the signal $z(t)$. Even though $h(t)$ does not have compact support, it can be well approximated by a truncated Gaussian, which does have compact

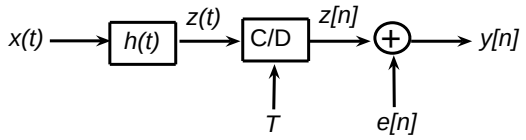


Figure 1: Block diagram showing our problem setup. $x(t)$ is a signal with FRI given by (2) and $h(t)$ is the Gaussian filter with width σ_h given by (3). $e[n]$ is i.i.d. Gaussian noise with standard deviation σ_e and $y[n]$ are the noisy samples. From $y[n]$ we will estimate the parameters that describe $x(t)$, namely $\{(c_k, t_k)\}_{k=1}^K$, and σ_e, the standard deviation of the noise.

support. The filtered signal $z(t)$ is sampled at rate of $1/T$ Hz to obtain $z[n] = z(nT)$ for $n = 0, 1, \ldots, N-1$. Finally, additive white Gaussian noise (AWGN) $e[n]$ is added to $z[n]$ to give $y[n]$. Therefore, the whole acquisition process from $x(t)$ to $\{y[n]\}_{n=0}^{N-1}$ can be represented by the model $\mathcal{M}$

$$\mathcal{M}: \quad y[n] = \sum_{k=1}^{K} c_k \exp\left(-\frac{(nT - t_k)^2}{2\sigma_h^2}\right) + e[n] \quad (4)$$

for $n = 0, 1, \ldots, N-1$. The amount of noise added is a function of σ_e. We define the signal-to-noise ratio (SNR) in dB as

$$\mathrm{SNR} \triangleq 10 \log_{10}\left(\frac{\sum_{n=0}^{N-1} |z[n]|^2}{\sum_{n=0}^{N-1} |z[n] - y[n]|^2}\right) \text{ dB}. \quad (5)$$

In the sequel, we will use boldface to denote vectors. In particular,

$$\mathbf{y} = [y[0], y[1], \ldots, y[N-1]]^\top, \quad (6)$$
$$\mathbf{c} = [c_1, c_2, \ldots, c_K]^\top, \quad (7)$$
$$\mathbf{t} = [t_1, t_2, \ldots, t_K]^\top. \quad (8)$$

We will be measuring the performance of our reconstruction algorithms by using the normalized reconstruction error

$$\mathcal{E} \triangleq \frac{\int_{-\infty}^{\infty} |z_{\mathrm{est}}(t) - z(t)|^2 \, dt}{\int_{-\infty}^{\infty} |z(t)|^2 \, dt}, \quad (9)$$

where $z_{\mathrm{est}}(t)$ is the reconstructed version of $z(t)$. By construction $\mathcal{E} \geq 0$ and the closer $\mathcal{E}$ is to 0, the better the reconstruction algorithm. The problem can be summarized as: *Given* $\mathbf{y} = \{y[n] \,|\, n = 0, \ldots, N-1\}$ *and the model* $\mathcal{M}$*, estimate the parameters* $\{(c_k, t_k)\}_{k=1}^K$*. Also estimate the noise variance* σ_e^2.

Ideally, we would like to minimize $\mathcal{E}$ in (9) directly, but this does not seem to be tractable since the dependence of $y[n]$ on $\{t_k\}_{k=1}^K$ is highly nonlinear. Thus, we propose the use of a stochastic algorithm (known as the Gibbs sampler) for the maximum likelihood (ML) estimation of $\{t_k\}_{k=1}^K$. The Gibbs sampler is a proxy for minimizing $\mathcal{E}$. This is followed by linear least squared error (LLSE) estimation of $\{c_k\}_{k=1}^K$ as a tractable and effective means for approximate minimization of $\mathcal{E}$.

3. Presentation of the Gibbs Sampler

The algorithm introduced in [7] is a stochastic optimization procedure based on Gibbs sampling to estimate $\boldsymbol{\theta} = \{\mathbf{c}, \mathbf{t}, \sigma_e\}$. Detailed derivations and a self-contained introduction to Gibbs sampling are given in [7], and code written in MATLAB can be found at http://web.mit.edu/∼vtan/frimcmc. Here, we merely summarize the main steps of the algorithm and the intuition behind Gibbs sampling.

The overall procedure is given in Algorithm 1. The algorithm uses Gibbs sampling (Algorithm 2) to estimate the set of Dirac positions $\{t_k\}_{k=1}^K$. It then uses a least-squares procedure to estimate the weights $\{c_k\}_{k=1}^K$. The basic idea of Gibbs sampling is to exploit the fact that it is easier to compute samples drawn approximately according to the posterior distribution of the parameters given the data than it is to directly minimize $\mathcal{E}$. This is true when one can analytically determine the conditional distribution of one parameter given the remaining parameters and the data. (The required derivations are presented in [7].) After a number of iterations I_b called the *burn-in period*, samples drawn through Gibbs sampling can be treated as if they are drawn from the true posterior. Thus, samples drawn in I additional iterations can be averaged to obtain a good approximation of the mean of the posterior distribution.

Algorithm 1 Parameter Estimation and Signal Reconstruction Algorithm

Require: Data $\mathbf{y}$, Model $\mathcal{M}$

1: Obtain estimates $\{\hat{t}_k\}_{k=1}^K$ and $\hat{\sigma}_e$ using the Gibbs sampler detailed in Algorithm 2 with the data $\mathbf{y}$ and the model $\mathcal{M}$ given in (4).
2: Obtain estimates $\{\hat{c}_k\}_{k=1}^K$ using a linear least squares estimation procedure and $\{\hat{t}_k\}_{k=1}^K$ from the Gibbs sampler.
3: Compute $z_{\mathrm{est}}(t) = \hat{x}(t) * h(t)$ given the parameters $\{(\hat{c}_k, \hat{t}_k)\}_{k=1}^K$ and the known pulse $h(t)$.
4: Compute reconstruction error $\mathcal{E}$ given in (9).

Algorithm 2 The Gibbs Sampling Algorithm

Require: $\mathbf{y}, I, I_b, \boldsymbol{\theta}^{(0)} = \{\mathbf{c}^{(0)}, \mathbf{t}^{(0)}, \sigma_e^{(0)}\}$

1: **for** $i \leftarrow 1 : I + I_b$ **do**
2: $\quad c_1^{(i)} \sim p(c_1 | c_2^{(i-1)}, c_3^{(i-1)}, \ldots, c_K^{(i-1)}, \mathbf{t}^{(i-1)} \sigma_e^{(i-1)})$
3: $\quad c_2^{(i)} \sim p(c_2 | c_1^{(i)}, c_3^{(i-1)}, \ldots, c_K^{(i-1)}, \mathbf{t}^{(i-1)} \sigma_e^{(i-1)})$
4: $\quad \vdots \sim \vdots$
5: $\quad c_K^{(i)} \sim p(c_K | c_1^{(i)}, c_2^{(i)}, \ldots, c_{K-1}^{(i)}, \mathbf{t}^{(i-1)}, \sigma_e^{(i-1)})$
6: $\quad t_1^{(i)} \sim p(t_1 | \mathbf{c}^{(i)}, t_2^{(i-1)}, t_3^{(i-1)}, \ldots, t_K^{(i-1)}, \sigma_e^{(i-1)})$
7: $\quad t_2^{(i)} \sim p(t_2 | \mathbf{c}^{(i)}, t_1^{(i)}, t_3^{(i-1)}, \ldots, t_K^{(i-1)}, \sigma_e^{(i-1)})$
8: $\quad \vdots \sim \vdots$
9: $\quad t_K^{(i)} \sim p(t_K | \mathbf{c}^{(i)}, t_1^{(i)}, t_2^{(i)}, \ldots, t_{K-1}^{(i)}, \sigma_e^{(i-1)})$
10: $\quad \sigma_e^{(i)} \sim p(\sigma_e | \mathbf{c}^{(i)}, \mathbf{t}^{(i)})$
11: **end for**
12: Compute $\hat{\boldsymbol{\theta}}_{\mathrm{MMSE}}$ using least squares
13: **return** $\hat{\boldsymbol{\theta}}_{\mathrm{MMSE}}$

Sampling c_k. c_k is sampled from a Gaussian distribution given by

$$p(c_k|\boldsymbol{\theta}_{-c_k}, \mathbf{y}, \mathcal{M}) = \mathcal{N}\left(c_k; -\frac{\beta_k}{2\alpha_k}, \frac{1}{2\alpha_k}\right), \quad (10)$$

where

$$\alpha_k \triangleq \frac{1}{2\sigma_e^2} \sum_{n=0}^{N-1} \exp\left(-\frac{(nT - t_k)^2}{\sigma_h^2}\right), \quad (11)$$

$$\beta_k \triangleq \frac{1}{\sigma_e^2} \sum_{n=0}^{N-1} \exp\left(-\frac{(nT - t_k)^2}{2\sigma_h^2}\right) \times \left\{ \sum_{\substack{k'=1 \\ k' \neq k}}^{K} c_{k'} \exp\left(-\frac{(nT - t_{k'})^2}{2\sigma_h^2}\right) - y[n] \right\}. \quad (12)$$

It is easy to sample from Gaussian densities when the parameters (α_k, β_k) have been determined.

Sampling t_k. t_k is sampled from a distribution of the form

$$p(t_k|\boldsymbol{\theta}_{-t_k}, \mathbf{y}, \mathcal{M}) \propto \exp\left[-\frac{1}{2\sigma_e^2} \sum_{n=0}^{N-1} \gamma_k \times \exp\left(-\frac{(nT - t_k)^2}{\sigma_h^2}\right) + \nu_k \exp\left(-\frac{(nT - t_k)^2}{2\sigma_h^2}\right) \right] \quad (13)$$

where

$$\gamma_k \triangleq c_k^2, \quad (14)$$

$$\nu_k \triangleq 2c_k \left\{ \sum_{\substack{k'=1 \\ k' \neq k}}^{K} c_{k'} \exp\left(-\frac{(nT - t_{k'})^2}{2\sigma_h^2}\right) - y[n] \right\}. \quad (15)$$

It is not straightforward to sample from this distribution. We used rejection sampling [5, 8] to generate samples $t_k^{(i)}$ from $p(t_k|\boldsymbol{\theta}_{-t_k}, \mathbf{y}, \mathcal{M})$. The proposal distribution $\tilde{q}(t_k)$ was chosen to be an appropriately scaled Gaussian, since it is easy to sample from Gaussians.

Sampling σ_e. σ_e is sampled from the 'Square-root Inverted-Gamma' [1] distribution $\mathcal{IG}^{-1/2}(\sigma_e; \varphi, \lambda)$,

$$p(\sigma_e|\boldsymbol{\theta}_{-\sigma_e}, \mathbf{y}, \mathcal{M}) = \frac{2\lambda^{\varphi}\sigma_e^{-(2\varphi+1)}}{\Gamma(\varphi)} \exp\left(-\frac{\lambda}{\sigma_e^2}\right) \mathbb{I}_{[0,+\infty)}(\sigma_e), \quad (16)$$

where

$$\varphi \triangleq \frac{N}{2}, \quad (17)$$

$$\lambda \triangleq \frac{1}{2}\left[y[n] - \sum_{k=1}^{K} c_k \exp\left(-\frac{(nT - t_k)^2}{2\sigma_h^2}\right) \right]^2 \quad (18)$$

Thus the distribution of the variance of the noise σ_e^2 is Inverted Gamma, which corresponds to the conjugate prior of σ_e^2 in the expression of $\mathcal{N}(e; 0, \sigma_e^2)$ [1] and thus it is easy to sample from.

	K	N	σ_e	SNR
AF/RF (Fig. 2(a))	5	30	10^{-6}	137 dB
GS (Fig. 2(b))	5	30	2.5	10.2 dB

Table 1: Parameter values for comparing annihilating filter and root-finding (AF/RF) against Gibbs sampling (GS).

4. Numerical Results and Experiments

In this section, the annihilating filter and root-finding algorithm [10] provides a baseline for comparison. After exhibiting its instability, we provide simulation results to validate the accuracy of the algorithm we proposed in Section 3. More extensive experimentation, including comparisons to [3] and applications to an audio signal, is reported in [7].

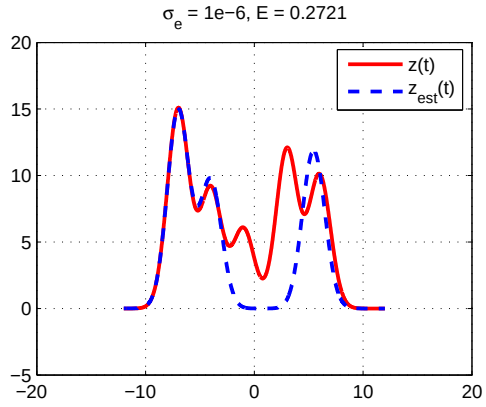
4.1 Annihilating Filter and Root-Finding

In [10], for signals of the form (2) and certain sampling kernels, the annihilating filter was used as a means to locate the t_k values. Subsequently a least squares approach yielded the weights c_k. It was shown that in the noiseless scenario, this method recovers the parameters exactly. In the same paper, a method for dealing with noisy samples is suggested. Unfortunately, this method seems to be inherently ill-conditioned. In Fig. 2, we show a pair of simulations with the parameters as tabulated in Table 1. We observe from Fig. 2(a) that (even with an oversampling factor of $N/(2K) = 3$) the annihilating filter and root-finding method is not robust to even a miniscule amount of added noise.

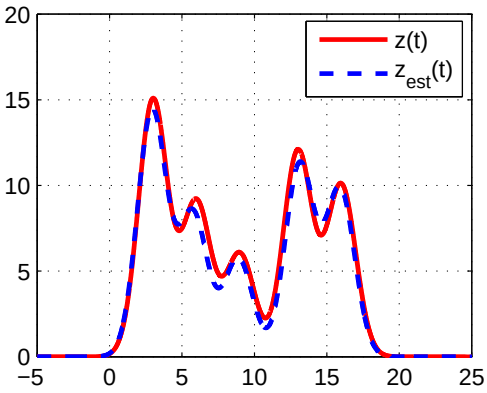
4.2 Gibbs Sampling Algorithm

Initial Demonstration. To demonstrate the evolution the Gibbs sampler, we performed an initial experiment with parameters as above, with the exception that the noise standard deviation was increased to $\sigma_e = 2.5$, giving an SNR of 10.2 dB. We plot the iterates of the most challenging parameters—the t_ks—in Fig. 3. We observe that the sampler converges in fewer than 20 iterations for this run, even though the parameter values were initialized far from their optimal values. The true filtered signal $z(t)$ and its estimate $z_{\text{est}}(t)$ are plotted in Fig. 2(b). Note the close similarity between $z(t)$ and $z_{\text{est}}(t)$.

Further Experiments on Simulated Data. To further validate our algorithm, we performed extensive simulations on different problem sizes with different levels of noise [7]. These experiments support the conclusion that the Gibbs sampler algorithm is insensitive to initialization. It *always* finds approximately optimal estimates from any starting point because the Markov chain provably converges to the stationary distribution [8]. We also find that the noise standard deviation σ_e can be estimated accurately; this may be important in some applications.



(a) The reconstruction using annihilating filter and root-finding completely breaks down when noise of a small standard deviation $\sigma_e = 10^{-6}$ (SNR = 137 dB) is added.



(b) The Gibbs sampling technique gives a much better reconstruction even at a higher noise level $\sigma_e = 2.5$ (SNR = 10.2 dB).

Figure 2: Demonstration of the instability of annihilating filter/root-finding approach and the improvement from Gibbs sampling.

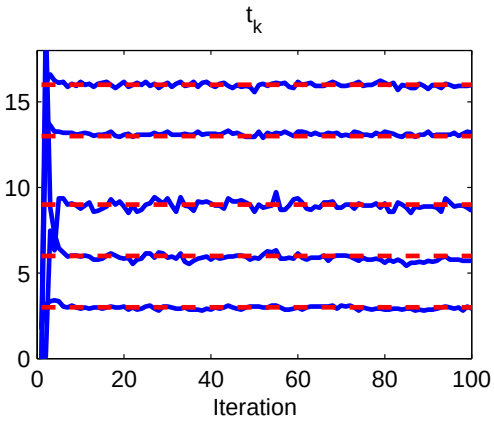


Figure 3: Evolution of the t_ks in the GS algorithm. The true values are indicated by the broken red lines.

5. Concluding Comments

We addressed the problem of reconstructing a signal with FRI given noisy samples. We showed that it is possible to circumvent some of the problems of the annihilating filter and root-finding approach [3, 10]. We introduced the Gibbs sampling algorithm to find the locations and augmented with a least squares approach to find the weights. The success of the Gibbs sampling algorithm does not depend on the choice of kernel $h(t)$, but rather the i.i.d. Gaussian noise assumption. The formulation of the Gibbs sampler does not depend on the specific form of $h(t)$. In fact, we used a Gaussian sampling kernel to illustrate that our algorithm is not restricted to the classes of kernels considered in [2].

A natural extension to our work here is to assign structured priors to $\mathbf{c}$, $\mathbf{t}$ and σ_e. These priors can themselves be dependent on their own set of *hyperparameters*, giving a hierarchical Bayesian formulation. In this way, there would be greater flexibility in the parameter estimation process. We can also seek to improve on the computational load of the algorithms introduced here and in particular the sampling of t_k via rejection sampling.

References:

[1] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. Wiley, 1st edition, 2001.

[2] P. L. Dragotti, M. Vetterli, and T. Blu. Sampling moments and reconstructing signals of finite rate of innovation: Shannon meets Strang–Fix. *IEEE Trans. Signal Processing*, 55(5):1741–1757, 2007.

[3] I. Maravic and M. Vetterli. Sampling and reconstruction of signals with finite rate of innovation in the presence of noise. *IEEE Trans. Signal Processing*, 53(8):2788–2805, 2005.

[4] H. Nyquist. Certain topics in telegraph transmission theory. *Trans. American Institute of Electrical Engineers*, 47:617–644, April 1928.

[5] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. New York: Springer-Verlag, 2nd edition, 2004.

[6] C. E. Shannon. Communication in the presence of noise. *Proc. Institute of Radio Engineers*, 37(1):10–21, January 1949.

[7] V. F. Y. Tan and V. K. Goyal. Estimating signals with finite rate of innovation from noisy samples: A stochastic algorithm. *IEEE Trans. Signal Process.*, 56(10):5135–5146, October 2008.

[8] L. Tierney. Markov chains for exploring posterior distributions. Technical Report 560, School of Statistics, Univ. of Minnesota, March 1994.

[9] M. Unser. Sampling–50 years after Shannon. *Proc. IEEE*, 88(4):569–587, 2000.

[10] M. Vetterli, P. Marziliano, and T. Blu. Sampling signals with finite rate of innovation. *IEEE Trans. Signal Processing*, 50(6):1417–1428, 2002.

[11] R. H. Walden. Analog-to-digital converter survey and analysis. *IEEE J. Selected Areas of Communication*, 17(4):539–550, April 1999.

The Generalized Annihilation Property

A Tool For Solving Finite Rate of Innovation Problems

Thierry Blu

The Chinese University of Hong Kong, Shatin N.T., Hong Kong
thierry.blu@m4x.org

Abstract:

We describe a property satisfied by a class of nonlinear systems of equations that are of the form $\mathbf{F}(\Omega)\mathrm{X} = \mathrm{Y}$. Here $\mathbf{F}(\Omega)$ is a matrix that depends on an unknown K-dimensional vector Ω, X is an unknown K-dimensional vector and Y is a vector of $N \geq K$) given measurements. Such equations are encountered in superresolution or sparse signal recovery problems known as "Finite Rate of Innovation" signal reconstruction.

We show how this property allows to solve explicitly for the unknowns Ω and X by a direct, non-iterative, algorithm that involves the resolution of two linear systems of equations and the extraction of the roots of a polynomial and give examples of problems where this type of solutions has been found useful.

1. Introduction

We consider the signal resulting from the convolution between a window $\varphi(t)$ and the sum of K Diracs with amplitude x_k located at time t_k. Given the N uniform samples y_n ($T =$ sampling step)

$$y_n = \sum_{k=1}^{K} x_k \varphi(nT - t_k) \qquad \text{where } n = 1, 2, \ldots, N, \quad (1)$$

then FRI problems (see [1, 2]) consist in retrieving the parameters t_k and x_k. Solving such problems is conceptually interesting because it shows how to break the standard Nyquist-Shannon bandlimitation rule for the exact reconstruction of signals from their uniform samples [3].

The system of consistent equations (1) can be expressed under the generic form of a nonlinear problem as shown in Fig. 1 (see next page), where the parameters $\Omega = [\omega_1, \omega_2, \ldots \omega_K]$ are related unambiguously to the unknowns t_k's. Because of the variety of settings adapted to this general approach, it happens to be necessary to distinguish between the parameters ω_k—which we shall call "abstract parameters"—and the locations t_k: typically, the ω_k's will be the zeros of some polynomial and from these ω_k's, we will be able to retrieve the t_k's using a functional relation of the form $\omega_k = \lambda(t_k)$ for some invertible function $\lambda(t)$.

At first sight, solving such a nonlinear system of equations is a daunting task. Fortunately, if the matrix $\mathbf{F}(\Omega)$ satisfies a property that we shall call "Generalized Annihilation Property" (GAP), this reduces to solving two linear systems of equations sandwiching a nonlinear step that amounts to polynomial root extraction in practical cases. The filters $\varphi(t)$ that satisfy the GAP are thus especially interesting, since the related FRI problems enjoy a straight non-iterative solution.

2. The Generalized Annihilation Property (GAP)

We carry on with the previously identified general nonlinear problem, namely

$$\mathbf{F}(\Omega)\,\mathrm{X} = \mathrm{Y}, \quad (3)$$

where the unknowns are $\Omega = [\omega_1, \omega_2, \ldots \omega_K]$ and $\mathrm{X} = [x_1, x_2, \ldots x_K]$, and where the measurements are $\mathrm{Y} = [y_1, y_2, \ldots y_N]$.

This system is said to satisfy the Generalized Annihilation Property whenever there exist $K + 1$ constant matrices, $\mathbf{A}_k$, and $K + 1$ scalar functions of Ω, $h_k(\Omega)$, such that we have the identity

$$\sum_{k=0}^{K} h_k(\Omega)\,\mathbf{A}_k\,\mathbf{F}(\Omega) = \mathbf{0}. \quad (4)$$

for any vector of parameters Ω. By right multiplying with X, the above equation implies that any solution Ω of (3) is also a solution of the (generalized) annihilation equation

$$\sum_{k=0}^{K} h_k(\Omega)\mathbf{A}_k\mathrm{Y} = 0. \quad (5)$$

This equation can be expressed in a matrix form $\mathbf{A}\mathrm{H} = 0$ where the unknown is $\mathrm{H} = [\,h_0(\Omega), h_1(\Omega), \ldots, h_K(\Omega)\,]^{\mathrm{T}}$ and the matrix $\mathbf{A} = \big[\,\mathbf{A}_0\mathrm{Y}, \mathbf{A}_1\mathrm{Y}, \ldots, \mathbf{A}_K\mathrm{Y}\,\big]$. Thus, in order to solve (3) for Ω and X, the idea consists in finding the scalar coefficients $h_k(\Omega)$ that satisfy (5), then retrieving $\omega_1, \omega_2, \ldots, \omega_K$ from the knowledge of $h_k(\Omega)$, and finally finding X such that $\mathbf{F}(\Omega)\,\mathrm{X} = \mathrm{Y}$. Without elaborating on the conditions that make this solution unique, a

$$\underbrace{\begin{bmatrix} \varphi(T-t_1) & \varphi(T-t_2) & \cdots & \varphi(T-t_K) \\ \varphi(2T-t_1) & \varphi(2T-t_2) & \cdots & \varphi(2T-t_K) \\ \vdots & \vdots & & \vdots \\ \varphi(NT-t_1) & \varphi(NT-t_2) & \cdots & \varphi(NT-t_K) \end{bmatrix}}_{\mathbf{F}(\Omega)} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_K \end{bmatrix}}_{\mathrm{X}} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{\mathrm{Y}} \tag{2}$$

Figure 1: Algebraic equivalent of the consistency equations (1).

minimal requirement is that the matrices $\mathbf{A}_k$ have at least K rows.

In the simple case where the $h_k(\Omega)$'s are related to the ω_k's through a polynomial relation

$$\sum_{k=0}^{K} h_k(\Omega) z^{-k} = \prod_{k=1}^{K} (1 - \omega_k z^{-1}), \tag{6}$$

solving (3) boils down to a three-step algorithm that can be summarized as follows:

1. Compute a solution $\mathrm{H} = [\,1, h_1, \ldots, h_{K-1}, h_K\,]^{\mathrm{T}}$ of
$$\big[\,\mathbf{A}_0\mathrm{Y}, \mathbf{A}_1\mathrm{Y}, \ldots, \mathbf{A}_K\mathrm{Y}\,\big]\,\mathrm{H} = 0;$$
2. Compute the roots ω_k of the z-transform $H(z) = \sum_{k=0}^{K} h_k z^{-k}$;
3. Compute a solution X of $\mathbf{F}(\Omega)\,\mathrm{X} = \mathrm{Y}$.

Example—Spectral estimation problems boil down to a nonlinear problem of the form (3) involving the *Vandermonde* matrix:

$$\mathbf{F}(\Omega) = \begin{bmatrix} \omega_1 & \omega_2 & \cdots & \omega_K \\ \omega_1^2 & \omega_2^2 & \cdots & \omega_K^2 \\ \vdots & \vdots & & \\ \omega_1^N & \omega_2^N & \cdots & \omega_K^N \end{bmatrix}$$

where the frequencies to retrieve, f_k, are related to ω_k through $\omega_k = e^{j2\pi f_k}$. This problem satisfies the GAP for band-diagonal matrices $\mathbf{A}_k$ which are more precisely given by:

$$\mathbf{A}_k = \big[\,\mathbf{0}_{N-K,k}\;\; \mathbf{I}_{N-K}\;\; \mathbf{0}_{N-K,K-k}\,\big],$$

where $\mathbf{0}_{m,n}$ is the $m \times n$ zero matrix and $\mathbf{I}_n$ is the $n \times n$ identity matrix. A minimal—yet not sufficient—condition for the unicity of the solution is $N \geq 2K$. Since the $\mathbf{A}_k$ can be seen as shifting operators by k samples, the annihilation equation is analogous to a filtering equation—with an annihilating filter. The annihilation algorithm is then equivalent to Prony's method [4]. Of course, spectral estimation in the presence of noise has been addressed by numerous researchers since the 1970's [5, 6, 7, 8, 9, 10, 11].

3. Some GAP Kernels

The GAP is actually shared by many interesting filters that can be used in sampling schemes, resulting in easily solvable FRI problems. Among them, the first ones to be identified were the periodized sinc, the infinite (i-e., not periodized) sinc and the Gaussian kernels [1]. Even more interestingly, recent research indicates that this property may somewhat be related to the Strang-Fix conditions which makes a very intriguing connection with approximation theory [12], and considerably broadens the class of FRI-admissible kernels. In all cases investigated so far, the scalar coefficients $h_k(\Omega)$ satisfy (6).

3.1 Periodized sinc (Dirichlet) filter

Solving the FRI problem in the case of a periodic stream of Diracs is equivalent to considering (1) where φ is a periodized sinc kernel, e.g., a Dirichlet kernel

$$\varphi(t) = \sum_{k' \in \mathbb{Z}} \mathrm{sinc}(B(t - k'\tau)) = \frac{\sin(\pi B t)}{B\tau \sin(\pi t/\tau)}$$

where τ is the period of the Dirac stream and B some bandwith (chosen so that $B\tau$ is an odd integer) [2]. This problem can be reformulated using the annihilation equation (4) by defining the following annihilation matrices

$$\mathbf{A}_k = \big[\,\mathbf{0}_{B\tau-K,k}\;\; \mathbf{I}_{B\tau-K}\;\; \mathbf{0}_{B\tau-K,B\tau-k}\,\big]\,\mathbf{W}$$

where $\mathbf{W} = [e^{-j2\pi mn/N}]$ for $|m| \leq \lfloor B\tau/2 \rfloor$ and $1 \leq n \leq N$, is the N-DFT submatrix of size $B\tau \times N$. Then, the abstract parameters ω_k are related to the locations t_k through $\omega_k = e^{-j2\pi t_k/\tau}$. This kernel has been found useful for the estimation of UWB channels [13] and for image superresolution [14].

3.2 Infinite sinc filter

The filter $\varphi(t)$ is given by $\varphi(t) = \mathrm{sinc}\, Bt$ with $B = 1/T$. When $\big(\varphi * x\big)(t)$ is sampled uniformly at frequency B, the nonlinear system of equations satisfies the GAP. The abstract parameters ω_k are related to the locations t_k through

$\omega_k = t_k$ and the annihilation matrices are given by

$$\mathbf{A}_k = \begin{bmatrix} \binom{K}{K} & \binom{K}{K-1} & \cdots & \binom{K}{0} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & & \ddots & 0 \\ 0 & \cdots & \cdots & \binom{K}{K} & \binom{K}{K-1} & \cdots & \binom{K}{0} \end{bmatrix} \times \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & 2^k & \ddots & & \vdots \\ \vdots & \ddots & 3^k & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & N^k \end{bmatrix}$$

3.3 Gaussian filter

The filter $\varphi(t)$ is given by $\varphi(t) = \exp(-t^2/\sigma^2)$. When $\big(\varphi * x\big)(t)$ is sampled uniformly at frequency T^{-1}, the nonlinear system of equations satisfies the GAP. The abstract parameters ω_k are related to the locations t_k through $\omega_k = \exp(2t_kT/\sigma^2)$ and the annihilation matrices are given by

$$\mathbf{A}_k = \begin{bmatrix} \mathbf{0}_{N-K,k} & \mathbf{I}_{N-K} & \mathbf{0}_{N-K,K-k} \end{bmatrix} \times \begin{bmatrix} e^{\frac{T^2}{\sigma^2}} & 0 & \cdots & \cdots & 0 \\ 0 & e^{\frac{(2T)^2}{\sigma^2}} & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & e^{\frac{(NT)^2}{\sigma^2}} \end{bmatrix}$$

A version of this solution (actually, for a Gabor kernel) was used in Optical Coherence Tomography, showing the possibility to resolve slices of a microscopic sample below the coherence length of the illuminating reference light [15].

3.4 Finite Support Strang-Fix filters

Through linear combinations of its shifts, the finite support filter $\varphi(t)$ is assumed to reconstruct polynomials up to some degree $L-1$ (standard Strang-Fix condition [16]) or exponentials $e^{a_l t}$ where $a_l - a_0$ is linear with $l = 0, 1, \ldots, L-1$. More precisely, in the standard Strang-Fix case, we denote by $c_{l,n}$ the coefficients such that

$$\sum_{n\in\mathbb{Z}} c_{l,n}\varphi(nT - t) = t^l \quad \text{where } l = 0, 1, \ldots, L-1,$$

by T the sampling step, and by $[0, S]$ the support of $\varphi(t)$. Then, the abstract parameters ω_k are related to the locations t_k through $\omega_k = t_k$ and the annihilation matrices are given by

$$\mathbf{A}_k = \begin{bmatrix} c_{k,1} & c_{k,2} & \cdots & c_{k,N} \\ c_{k-1,1} & c_{k-1,2} & \cdots & c_{k-1,N} \\ \vdots & \vdots & & \vdots \\ c_{k-L+1,1} & c_{k-L+1,2} & \cdots & c_{k-L+1,N} \end{bmatrix}.$$

Additionally, there is a constraint on the minimal number of samples N for the GAP to hold, which is that N be larger than $\lceil (S + \max_k\{t_k\})/T \rceil$.

4. Conclusion

We have shown how to unify the different techniques used in FRI signal reconstruction through an algebraic property that we call the Generalized Annihilation property. In essence, this property allows to solve nonlinear system of equations within two noniterative steps. We hope that this property can be used to solve other FRI problems (i.e, with new kernels) in particular in dimensions higher than 1 (for instance, like in [17]), and maybe to solve other types of problems not directly related to sampling.

References:

[1] M. Vetterli, P. Marziliano, and T. Blu, "Sampling signals with finite rate of innovation," *IEEE Transactions on Signal Processing*, vol. 50, pp. 1417–1428, June 2002.

[2] T. Blu, P.-L. Dragotti, M. Vetterli, P. Marziliano, and L. Coulot, "Sparse sampling of signal innovations," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 31–40, 2008.

[3] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423 and 623–656, July and October 1948.

[4] R. Prony, "Essai expérimental et analytique," *Annales de l'École Polytechnique*, vol. 1, no. 2, p. 24, 1795.

[5] P. Stoica and R. L. Moses, *Introduction to Spectral Analysis*. Upper Saddle River, NJ: Prentice Hall, 1997.

[6] S. M. Kay, *Modern Spectral Estimation—Theory and Application*. Englewood Cliffs, NJ: Prentice Hall, 1988.

[7] D. W. Tufts and R. Kumaresan, "Estimation of frequencies of multiple sinusoids: Making linear prediction perform like maximum likelihood," *Proceedings of the IEEE*, vol. 70, pp. 975–989, September 1982.

[8] S. M. Kay and S. L. Marple, "Spectrum analysis—a modern perspective," *Proc. IEEE*, vol. 69, pp. 1380–1419, November 1981.

[9] *Special Issue on Spectral Estimation, Proceedings of the IEEE*, vol. 70, September 1982.

[10] V. F. Pisarenko, "The retrieval of harmonics from a covariance function," *Geophysical Journal*, vol. 33, pp. 347–366, September 1973.

[11] R. Roy and T. Kailath, "ESPRIT–estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, pp. 984–995, July 1989.

[12] P.-L. Dragotti, M. Vetterli, and T. Blu, "Sampling moments and reconstructing signals of finite rate of innovation: Shannon meets Strang-Fix," *IEEE Transactions on Signal Processing*, vol. 55, pp. 1741–1757, May 2007. Part 1.

[13] I. Maravić, J. Kusuma, and M. Vetterli, "Low-sampling rate UWB channel characterization and synchronization," *Journal of Communications and Networks*, vol. 5, no. 4, pp. 319–327, 2003.

[14] L. Baboulaz and P.-L. Dragotti, "Exact feature extraction using finite rate of innovation principles with an application to image super-resolution," *IEEE Transactions on Image Processing*, vol. 18, pp. 281–298, February 2009.

[15] T. Blu, H. Bay, and M. Unser, "A new high-resolution processing method for the deconvolution of optical coherence tomography signals," in *Proceedings of the First IEEE International Symposium on Biomedical Imaging: Macro to Nano (ISBI'02)*, vol. III, (Washington DC, USA), pp. 777–780, July 7-10, 2002.

[16] G. Strang and G. Fix, "A Fourier analysis of the finite element variational method," in *Constructive Aspects of Functional Analysis* (G. Geymonat, ed.), pp. 793–840, Rome: Edizioni Cremonese, 1973.

[17] D. Kandaswamy, T. Blu, and D. Van De Ville, "Analytic sensing: reconstructing pointwise sources from boundary Laplace measurements," in *Proceedings of the SPIE Conference on Mathematical Imaging: Wavelet XII*, (San Diego CA, USA), August 26-August 30, 2007. To appear.

An "algebraic" reconstruction of piecewise-smooth functions from integral measurements

Dima Batenkov, Niv Sarig, Yosef Yomdin

Department of Mathematics, Weizmann institute of science, Rehovot, Israel.
{dima.batenkov, niv.sarig, yosef.yomdin}@weizmann.ac.il

1. Introduction

This paper presents some results on a well-known problem in Algebraic Signal Sampling and in other areas of applied mathematics: reconstruction of piecewise-smooth functions from their integral measurements (like moments, Fourier coefficients, Radon transform, etc.). Our results concern reconstruction (from the moments) of signals in two specific classes: linear combinations of shifts of a given function, and "piecewise D-finite functions" which satisfy on each continuity interval a linear differential equation with polynomial coefficients.

Let us start with some general remarks and a conjecture. It is well known that the error in the best approximation of a C^k-function f by an N-th degree Fourier polynomial is of order $\frac{C}{N^k}$. The same holds for algebraic polynomial approximation and for other basic approximation tools. However, for f with singularities, in particular, with discontinuities, the error is much larger: its order is only $\frac{C}{\sqrt{N}}$. Considering the so-called Kolmogorow N-width of families of signals with moving discontinuities one can show that *any linear approximation method provides the same order of error, if we do not fix a priori the discontinuities' position* (see [7], Theorem 2.10). Another manifestation of the same problem is the *"Gibbs effect" - a relatively strong oscillation of the approximating function near the discontinuities*. Practically important signals usually do have discontinuities, so the above feature of linear representation methods presents a serious problem in signal reconstruction. In particular, it visibly appears near the edges of images compressed by JPEG, as well as in the noise and low resolution of the CT and MRI images.

Recent non-linear reconstruction methods, in particular, Compressed Sensing ([2, 3]) and Algebraic Sampling ([4, 12, 14, 6, 9]), address this problem in many cases. Both approaches appeal to an a priori information on the character of the signals to be reconstructed, assuming their "simplicity" in one or another sense. Compressed sensing assumes only a sparse representation in a certain (wavelets) basis, and thus it presents a rather general and "universal" approach. Algebraic Sampling usually requires more specific a priori assumptions on the structure of the signals, but it promises a better reconstruction accuracy. In fact, we believe that ultimately the Algebraic Sampling approach has a potential to reconstruct "simple signals with singularities" as good as smooth ones. In particular, the results of [5, 11, 8, 17, 14] strongly support (also apparently do not accurately formulate and prove) the following conjecture:

There is a non-linear algebraic procedure reconstructing any signal in a class of piecewise C^k-functions (of one or several variables) from its first N Fourier coefficients, with the overall accuracy of order $\frac{C}{N^k}$. This includes the discontinuities' positions, as well as the smooth pieces over the continuity domains.

At present there are many approaches available to a robust detection of discontinuities from Fourier data (see [8, 5, 11] and references therein). The remaining problem seems to be an accurate estimate of the accuracy of the solution of the nonlinear systems arising. Our results below can be considered, in particular, as a step in this direction. On the other hand, they have been motivated by the results in [4, 12, 14], and in [9, 6].

2. Linear combinations of shifts of a given function

Reconstruction of this class of signals from sampling has been described in [4, 12]. We study a rather similar problem of reconstruction from the moments. Our method is based on the following approach: we construct convolution kernels dual to the monomials. Applying these kernels, we get a Prony-type system of equations on the shifts and amplitudes.

Let us restate a general reconstruction problem, as it appears in our specific setting. We want to reconstruct signals of the form

$$F(x) = \sum_{i=1}^{N} \sum_{j,l} a_{i,j,l} f_i^{(l)}(x + x^j) \qquad (1)$$

where the f_i's are known functions of $x = (x_1, \ldots, x_d)$, and the form (1) of the signal is known a priori. The parameters $a_{i,j,l}$, $x^j = (x_1^j, \ldots, x_d^j)$ are to be found from a finite number of "measurements", i.e. of linear (usually integral) functionals like polynomial moments, Fourier moments, shifted kernels, evaluation over some grid and more.

In this paper we consider only linear combinations of shifts of one known function f (although the method of "convolution dual" can be extended to several shifted functions and their derivatives - see [16]). First we consider general integral "measurements" and then restrict

ourselves to the moments and Fourier coefficients. In what follows $x = (x_1, \dots, x_d), t = (t_1, \dots, t_d)$, j is a scalar index, while $k = (k_1, \dots, k_d)$, $i = (i_1, \dots, i_d)$ and $n = (n_1, \dots, n_d)$ are multi-indices. Partial ordering of multi-indices is given by $k \le k' \Leftrightarrow k_p \le k'_p,\ p = 1, \dots, d$. So we have

$$F(x) = \sum_{j=1}^{s} a_j f(x + x^j). \tag{2}$$

Let the measurements $\mu_k(F)$ be given by $\mu_k(F) = \int F(t)\varphi_k(t)dt$, for a certain (multi)-sequence of functions $\varphi_k(t),\ k \ge 0 = (0, \dots, 0)$.
Given f and $\varphi = \{\varphi_k(t)\},\ k \ge 0$ we now try to find certain "triangular" linear combinations

$$\psi_k(t) = \sum_{0 \le i \le k} C_{i,k}\varphi_i(t) \tag{3}$$

forming, in a sense, some "f-convolution dual" functions (similar to a bi-orthogonal set of function) with respect to the system $\varphi_k(t)$. More accurately, we require that

$$\int f(t + x)\psi_k(t) = \varphi_k(x). \tag{4}$$

We shall call a sequence $\psi = \{\psi_k(t)\}$ satisfying (3), (4) f - convolution dual to φ. Below we find convolution dual systems to the usual and exponential monomials.
We consider a general problem of finding convolution dual sequences to a given sequence of measurements as an important step in the reconstruction problem. Notice that it can be generalized by dropping the requirement of a specific representation (3): $\psi_k(t) = \sum_{i=0}^{k} C_{i,k}\varphi_i(t)$. Instead we can require only that $\int f(t)\psi_k(t)$ be expressible in terms of the measurements sequence μ_k. Also φ_k in (4) can be replaced by another a priori chosen sequence η_k. This problem leads, in particular, to certain functional equations, satisfied by polynomials and exponents (as well as exponential polynomials and some kinds of elliptic functions).
Now we have the following result:

Theorem 1. *Let a sequence $\psi = \psi_k(t)$ be f-convolution dual to φ. Define M_k by $M_k = \sum_{0 \le i \le k} C_{i,k}\mu_i$. Then the parameters a_j and x^j in (2) satisfy the following system of equations ("generalized Prony system"):*

$$\sum_{j=1}^{s} a_j \varphi_k(x^j) = M_k,\ \ k = 0, \dots. \tag{5}$$

Proof We have $M_k = \sum_{0 \le i \le k} C_{i,k}\mu_i = \int F(t) \sum_{0 \le i \le k} C_{i,k}\varphi_i(t)dt = \int F(t)\psi_k(t) = \sum_{j=1}^{s} a_j \int f(t + x^j)\psi_k(t)dt = \sum_{j=1}^{s} a_j \varphi_k(x^j)$.
In specific examples we can find the minimal number of equations in (5) necessary to uniquely reconstruct the parameters a_j and x^j in (2).

2.1 Reconstruction from moments

We are given a finite number of moments of a signal F as in (2) in the form

$$m_n = \int F(t)t^n dt. \tag{6}$$

So here $\varphi_n(x) = x_1^{n_1} \cdots x_d^{n_d}$ for each multi-index $n = (n_1, \dots, n_d)$. We look for the dual functions ψ_n satisfying the convolution equation

$$\int f(t + x)\psi_n(t)dt = x^n \tag{7}$$

for each multi-index n. To solve this equation we apply Fourier transform to both sides of (7). Assuming that $\hat{f}(\omega) \in C^\infty(\mathbb{R}^d), \hat{f}(0) \ne 0$ we find (see [16]) that there is a unique solution to (7) provided by

$$\varphi_n(x) = \sum_{k \le n} C_{n,k}x^k, \tag{8}$$

where

$$C_{n,k} = \frac{1}{(\sqrt{2\pi})^d} \binom{n}{k} (-i)^{n+k} \left[\frac{\partial^{n-k}}{\partial \omega^{n-k}} \Big|_{\omega=0} \frac{1}{\hat{f}(\omega)} \right].$$

This calculation is symbolic and works for more general cases. The actual calculation in our polynomial case is done using straightforward matrix calculations. We set the generalized polynomial moments as

$$M_n = \sum_{k \le n} C_{n,k}m_k \tag{9}$$

and obtain, as in Theorem 1, the following system of equations:

$$\sum_{j=1}^{s} a_j (x^j)^n = M_n,\ n \ge 0. \tag{10}$$

This system can be solved explicitly in a standard way (see, for example, [13, 4, 15]). In one-dimensional case it goes as follows (see [13]): from (10) we get that for $z = (z_1, \dots, z_d)$ the generalized moments generating function ($d = 1$ yet, notice that the formulas are still multi-dimensional)

$$I(z) = \sum_{n \in \mathbb{N}^d} M_n z^n = \sum_{j=1}^{s} a_j \prod_{l=1}^{d} \frac{1}{1 - x_l^j z_l} \tag{11}$$

is a rational function. Hence its Taylor coefficients satisfy linear recurrence relation, which can be reconstructed through a linear system with the Hankel-type matrix formed by an appropriate number of the moments M_n's. This is, essentially, a procedure of the diagonal Padé approximation for $I(z)$ (see [13]). The parameters a_j, x^j are finally reconstructed as the poles and the residues of $I(z)$. For several variables, although the formulas are the same as above, the generalization of the solution of the Prony system is more involved and should be addressed separately.
In one dimensional case with the derivatives $f^{(l)}$ included we have

$$F(x) = \sum_{j=1}^{s} \sum_{l=0}^{r} a_{j,l} f^{(l)}(x + x^j). \tag{12}$$

The corresponding moment-generating function in this case takes the form

$$I(z) = \sum_{j=1}^{s} \sum_{l=0}^{r} \sum_{q=0}^{l} \binom{l}{q} \frac{(-1)^{q+l} a_{j,l}/(x^j)^l}{(1 - x^j z)^{q+1}}. \tag{13}$$

which is still a rational function (d-dimensional case with derivatives is similar). We would like to stress that in this case the dual polynomials ψ_k are not changed and they are given as in (8). Therefore also the formula for the generalized moments M_n is the same as in (9).

2.2 Fourier case

In the same manner as in section 2.1 we now choose $\varphi_k(x) = e^{ikx}$. We get immediately $\psi_k(x) = \frac{1}{\hat{f}(k)} e^{-ikx}$. Indeed,

$$\int f(t+x)\psi_k(t)dt = \int f(t+x)\frac{1}{\hat{f}(k)}e^{ikt}dt =$$

$$\frac{\hat{f}(k)}{\hat{f}(k)}e^{-ikx} = \varphi_{-k}(x). \qquad (14)$$

Here the triangular system of equations (3) is actually not triangular any more but still since $\psi_k(x) = \frac{1}{\hat{f}(k)}\varphi_{-k}(x)$ we can express the generalized moments through the original ones via $M_k = \frac{1}{\hat{f}(k)}\mu_{-k}[F]$. Now exactly as before we can find a generalized Prony system in the form

$$\frac{1}{\hat{f}(k)}\mu_{-k}[F] = M_k = \sum_j a_j e^{-ikx_j} = \sum_j a_j \rho_j^k \quad (15)$$

where $\rho_j = e^{-ix_j}$. In this case we get a rational exponential generating function and we can find its poles and residues on the unit complex circle as we did in the polynomial case.

2.3 Further extensions

The approach above can be extended in the following directions: 1. Reconstruction of signals built from several functions or with the addition of dilations also can be investigated (a perturbation approach where the dilations are approximately 1 is studied in [15]). 2. Further study of "convolution duality" can significantly extend the class of signals and measurements allowing for a closed - form signal reconstruction.

3. Reconstruction of piecewise D-finite functions from moments

Let $g : [a, b] \to \mathbb{R}$ consist of $\mathcal{K}+1$ "pieces" $g_0, \dots g_{\mathcal{K}}$ with $\mathcal{K} \geq 0$ jump points

$$a = \xi_0 < \xi_1 \dots < \xi_{\mathcal{K}} < \xi_{\mathcal{K}+1} = b$$

Furthermore, let g satisfy on each continuity interval some linear homogeneous differential equation with polynomial coefficients: $\mathfrak{D}\, g_n \equiv 0,\ n = 0, \dots, \mathcal{K}$ where

$$\mathfrak{D} = \sum_{j=0}^{N}\left(\sum_{i=0}^{k_j} a_{i,j}x^i\right)\frac{d^j}{dx^j} \quad (a_{ij} \in \mathbb{R}) \qquad (16)$$

Each g_n may be therefore written as a linear combination of functions $\{u_i\}_{i=1}^N$ which are a basis for the space $\mathcal{N}_{\mathfrak{D}} = \{f : \mathfrak{D}\, f \equiv 0\}$:

$$g_n(x) = \sum_{i=1}^{N} \alpha_{i,n} u_i(x), \quad n = 0, 1, \dots, \mathcal{K} \qquad (17)$$

We term such functions g "piecewise D-finite". Many real-world signals may be represented as piecewise D-finite functions, in particular: polynomials, trigonometric functions, algebraic functions.
The sequence $\{m_k = m_k(g)\}$ is given by the usual moments

$$m_k(g) = \int_a^b x^k g(x)dx$$

We subsequently formulate the following

Piecewise D-finite Reconstruction Problem. *Given $N, \{k_i\}, \mathcal{K}, a, b$ and the moment sequence $\{m_k\}$ of a piecewise D-finite function g, reconstruct all the parameters $\{a_{i,j}\}, \{\xi_i\}, \{\alpha_{i,n}\}$.*

Below we state some results (see [1] for detailed proofs) which provide *explicit algebraic connections* between the above parameters and the measurements $\{m_k\}$.
The first two theorems assume a single continuity interval (compare [10]).

Theorem 2. *Let $\mathcal{K} = 0$ and $\mathfrak{D}\, g \equiv 0$ with $\mathfrak{D}$ given by* (16)*. Then the moment sequence $\{m_k(g)\}$ satisfies a linear recurrence relation*

$$\left((\mathrm{E} - a\,\mathrm{I})^N(\mathrm{E} - b\,\mathrm{I})^N \cdot \sum_{j=0}^{N}\sum_{i=0}^{k_j} a_{i,j}\Pi^{(i,j)}(k, \mathrm{E})\right)m_k = 0 \qquad (18)$$

where E *is the discrete forward shift operator and $\Pi^{(i,j)}(k, \mathrm{E})$ are monomials in* E *whose coefficients are polynomials in k: $\Pi^{(i,j)}(k, \mathrm{E}) = (-1)^j \frac{(i+k)!}{(i+k-j)!}\,\mathrm{E}^{i-j}$.*

Theorem 3. *Denote*

$$\mathcal{E}(\mathrm{E}) \stackrel{def}{=} (\mathrm{E} - a\,\mathrm{I})^N(\mathrm{E} - b\,\mathrm{I})^N, \qquad v_k^{(i,j)} \stackrel{def}{=} \big(\mathcal{E}(\mathrm{E}) \cdot \Pi^{(i,j)}(k, \mathrm{E})\big)m_k,$$

$$h_j(z) \stackrel{def}{=} \sum_{k=0}^{\infty} v_k^{(0,j)} z^k, \qquad G_j(x) \stackrel{def}{=} \mathcal{E}(x)\frac{d^j}{dx^j}g(x)$$

Assume the conditions of Theorem 2. Then

(1) The vector of the coefficients $\mathbf{a} = (a_{i,j})$ satisfies a linear homogeneous system

$$H\mathbf{a} = \begin{pmatrix} v_0^{(0,0)} & v_0^{(1,0)} & \dots & v_0^{(k_N,N)} \\ v_1^{(0,0)} & v_1^{(1,0)} & \dots & v_1^{(k_N,N)} \\ \vdots & \vdots & \vdots & \vdots \\ v_{\widehat{M}}^{(0,0)} & v_{\widehat{M}}^{(1,0)} & \dots & v_{\widehat{M}}^{(k_N,N)} \end{pmatrix}\begin{pmatrix} a_{0,0} \\ a_{1,0} \\ \vdots \\ a_{k_N,N} \end{pmatrix} = 0 \qquad (19)$$

for all $\widehat{M} \in \mathbb{N}$.

(2) $v_k^{(i,j)} = m_{i+k}\,(G_j(x))$. Consequently, $h_j(z)$ is the moment generating function of $G_j(x)$.

(3) Denote $p_j(x) = \sum_{i=0}^{k_j} a_{i,j}x^i$. Then the functions $\Phi = \{1, h_0(z), \dots h_N(z)\}$ are polynomially dependent: $\sum_{j=0}^{N} h_j(z)\big(z^{\max k_j}p_j(z^{-1})\big) = Q(z)$ where $Q(z)$ is a polynomial with $\deg Q < \max k_j$. The system of polynomials $\{z^{\max k_j}p_j(z^{-1})\}$ is called the Padé-Hermite form for Φ.

To handle the piecewise case, we represent the jump discontinuities by the step function $\mathcal{H}(x) \stackrel{\text{def}}{=} \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}$ and write g as a distribution

$$g(x) = \widetilde{g_0} + \sum_{n=1}^{\mathcal{K}} \widetilde{g_n}(x)\mathcal{H}(x - \xi_n) \tag{20}$$

Theorem 4. *Let $\mathcal{K} > 0$ and let g be as in* (20) *with operator $\mathfrak{D}$ annihilating every piece $\widetilde{g_n}$. Then the operator*

$$\widehat{\mathfrak{D}} \stackrel{def}{=} \left\{ \prod_{n=1}^{\mathcal{K}} (x - \xi_i)^N \, \mathrm{I} \right\} \cdot \mathfrak{D} \tag{21}$$

annihilates the entire g as a distribution. Consequently, conclusions of Theorems 2 and 3 hold with $\mathfrak{D}$ replaced by $\widehat{\mathfrak{D}}$ as in (21).

Proposition 5. *Let $\mathcal{K} \geq 0$ and $\{u_i\}_{i=1}^N$ be a basis for the space $\mathcal{N}_{\mathfrak{D}}$, where $\mathfrak{D}$ annihilates every piece of g. Assume* (17) *and denote $c_{i,k}^n = \int_{\xi_n}^{\xi_{n+1}} x^k u_i(x)$ for $n = 0, \ldots, \mathcal{K}$. A straightforward computation gives $\forall \widetilde{M} \in \mathbb{N}$:*

$$\begin{pmatrix} c_{1,0}^0 & \cdots & c_{N,0}^0 & \cdots & c_{N,0}^{\mathcal{K}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{1,\widetilde{M}}^0 & \cdots & c_{N,\widetilde{M}}^0 & \cdots & c_{N,\widetilde{M}}^{\mathcal{K}} \end{pmatrix} \begin{pmatrix} \alpha_{1,0} \\ \vdots \\ \alpha_{N,0} \\ \vdots \\ \alpha_{N,\mathcal{K}} \end{pmatrix} = \begin{pmatrix} m_0 \\ m_1 \\ \vdots \\ m_{\widetilde{M}} \end{pmatrix} \tag{22}$$

The above results can be combined as follows to provide a solution of the Reconstruction Problem:

(a) Let $N, \{k_i\}, \mathcal{K}, a, b$ and $\{m_k(g)\}$ be given. If $\mathcal{K} > 0$, replace $\mathfrak{D}$ (still unknown) with $\widehat{\mathfrak{D}}$ according to (21).

(b) Build the matrix H as in (19). Solve $H\mathbf{a} = 0$ and obtain the operator $\mathfrak{D}^* = \mathfrak{D}_{\mathbf{a}}$ which annihilates g.

(c) If $\mathcal{K} > 0$, factor out all the common roots of the polynomial coefficients of $\mathfrak{D}^*$ with multiplicity N. These are the locations of the jump points $\{\xi_n\}$. The remaining part is the operator $\mathfrak{D}^\dagger$ which annihilates every g_n.

(d) By now $\mathfrak{D}^\dagger$ and $\{\xi_n\}$ are known. So compute the basis for $\mathcal{N}_{\mathfrak{D}^\dagger}$ and solve (22).

The constants $\widehat{M}$ and $\widetilde{M}$ determine the minimal required size of the corresponding linear systems (19) and (22) in order for all the solutions of these systems to be also solutions of the original problem. It can be shown that:

1. There exists no uniform bound on $\widehat{M}$ without any additional information on the nature of the solutions. Explicit bounds may be obtained for simple function classes such as piecewise polynomials of bounded degrees or real algebraic functions.

2. For every specific $\mathfrak{D}$, an explicit bound $\widetilde{M} = \widetilde{M}(\mathfrak{D})$ may be computed for the system (22).

The above algorithm has been tested on exact reconstruction of piecewise polynomials, piecewise sinusoids and rational functions.

References:

[1] D.Batenkov, *Moment inversion problem for piecewise D-finite functions*, arXiv:0901.4665v2 [math.CA].

[2] E. J. Candeès. *Compressive sampling.* Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006. Vol. III, 1433–1452, Eur. Math. Soc., Zurich, 2006.

[3] D. Donoho, *Compressed sensing.* IEEE Trans. Inform. Theory 52 (2006), no. 4, 1289–1306.

[4] P.L. Dragotti, M. Vetterli and T. Blu, *Sampling Moments and Reconstructing Signals of Finite Rate of Innovation: Shannon Meets Strang-Fix*, IEEE Transactions on Signal Processing, Vol. 55, Nr. 5, Part 1, pp. 1741-1757, 2007.

[5] K. Eckhoff, *Accurate reconstructions of functions of finite regularity from truncated Fourier series expansions*, Math. Comp. 64 (1995), no. 210, 671–690.

[6] M. Elad, P. Milanfar, G. H. Golub, *Shape from moments—an estimation theory perspective*, IEEE Trans. Signal Process. **52** (2004), no. 7, 1814–1829.

[7] B. Ettinger, N. Sarig. Y. Yomdin, *Linear versus non-linear acqusition of step-functions*, J. of Geom. Analysis, 18 (2008), 2, 369-399.

[8] A. Gelb, E. Tadmor, *Detection of edges in spectral data II. Nonlinear enhancement*, SIAM J. Numer. Anal. 38 (2000), 1389-1408.

[9] B. Gustafsson, Ch. He, P. Milanfar, M. Putinar, *Reconstructing planar domains from their moments.* Inverse Problems 16 (2000), no. 4, 1053–1070.

[10] V. Kisunko, Cauchy type integrals and a D-moment problem. C.R. Math. Acad. Sci. Soc. R. Can. 29 (2007), no. 4, 115–122.

[11] G. Kvernadze, T. Hagstrom, H. Shapiro, *Locating discontinuities of a bounded function by the partial sums of its Fourier series.*, J. Sci. Comput. 14 (1999), no. 4, 301–327.

[12] I. Maravic and M. Vetterli, *Exact Sampling Results for Some Classes of Parametric Non-Bandlimited 2-D Signals*, IEEE Transactions on Signal Processing, Vol. 52, Nr. 1, pp. 175-189, 2004.

[13] E. M. Nikishin, V. N. Sorokin, *Rational Approximations and Orthogonality*, Translations of Mathematical Monographs, **Vol 92**, AMS, 1991.

[14] P. Prandoni, M. Vetterli, *Approximation and compression of piecewise smooth functions*, R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci. **357** (1999), no. 1760, 2573–2591.

[15] N. Sarig, Y. Yomdin, Signal Acquisition from Measurements via Non-Linear Models, C. R. Math. Rep. Acad. Sci. Canada Vol. 29 (4) (2007), 97-114.

[16] N. Sarig and Y. Yomdin, *Reconstruction of "Simple" Signals from Integral Measurements*, in preparation.

[17] E. Tadmor, High resolution methods for time dependent problems with piecewise smooth solutions. *Proceedings of the International Congress of Mathematicians*, **Vol. III** (Beijing, 2002), 747–757, Higher Ed. Press, Beijing, 2002.

Distributed Sensing of Signals Under a Sparse Filtering Model

Ali Hormati , Olivier Roy , Yue M. Lu and Martin Vetterli

Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland.

Abstract:

We consider the task of recovering correlated vectors at a central decoder based on fixed linear measurements obtained by distributed sensors. Two different scenarios are considered: In the case of universal reconstruction, we look for a sensing and recovery mechanism that works for *all* possible signals, whereas in the case of almost sure reconstruction, we allow to have a small set (with measure zero) of unrecoverable signals. We provide achievability bounds on the number of samples needed for both scenarios. The bounds show that *only* in the almost sure setup can we effectively exploit the signal correlations to achieve effective gains in sampling efficiency. In addition, we propose an efficient and robust distributed sensing and reconstruction algorithm based on annihilating filters.

1. Introduction

Consider two signals that are linked by an unknown filtering operation, where the filter is sparse in the time domain. Such models can be used, e.g., to describe the correlation between the transmitted and received signals in an unknown multi-path environment. We sample the two signals in a distributed setup: Each signal is observed by a different sensor, which sends a certain number of *non-adaptive* and *fixed* linear measurements of that signal to a central decoder. We study how the correlation induced by the above model can be exploited to reduce the number of measurements needed for perfect reconstruction at the central decoder, but *without* any inter-sensor communication during the sampling process.

Our setup is conceptually similar to the Slepian-Wolf problem in distributed source coding [6], which consists of correlated sources to be encoded separately and decoded jointly. While communication between the encoders is precluded, correlation between the measured data can be taken into account as an effective means to reduce the amount of information transmitted to the decoder. The main difference between our work and this classical distributed source coding setup is that we study a *sampling* problem and hence are only concerned about the number of sampling measurements we need to take, whereas the latter is about *coding* and hence uses bits as its "currency". From the sampling perspective, our work is closely related to the problem of distributed compressed sensing, first introduced in [1] (see also [4, 5]). In that framework, jointly sparse data need to be reconstructed based on linear projections computed by distributed sensors. In this paper, we first introduce in Section 2. a novel correlation model for distributed signals. Instead of imposing any sparsity assumption on the signals themselves (as in [1]), we assume that the signals are linked by some unknown sparse filtering operation. Such models can be useful in describing the signal correlation in several practical scenarios (e.g. multi-path propagation and binaural audio recoding). In Section 3., we introduce two strategies for the design of the sampling system: In the *universal* strategy, we seek to successfully sense and recover *all* signals, whereas in the *almost sure* strategy, we allow to have a small set (with measure zero) of unrecoverable signals. We establish the corresponding achievability bounds on the number of samples needed for the two strategies mentioned above. These bounds indicate that the sparsity of the filter can be useful only in the almost sure strategy. Since the algorithms that achieves the bounds are computationally prohibitive, we introduce in Section 4., a concrete distributed sampling and reconstruction scheme that can recover the original signals in an efficient and robust way. Finally, Section 5. presents an application of our results in the area of binaural hearing aids. A preliminary version of this work was also presented at ICASSP 2009. In this paper, we add results on the achievability bound for the almost sure setup as well as a new section on applications.

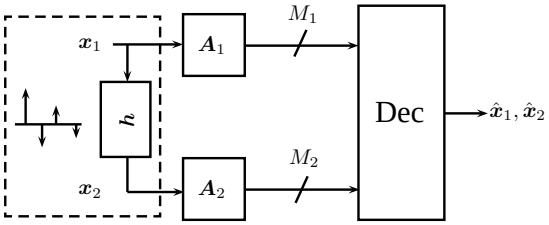


Figure 1: Distributed sensing setup. Signals $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are connected through an unknown sparse filter $\boldsymbol{h}$. The ith sensor ($i = 1, 2$) provides a M_i-dimensional observation of the signal $\boldsymbol{x}_i$ via a non-adaptive and fixed linear transform $\boldsymbol{A}_i$ to a central decoder.

2. The Correlation Model

Consider two signals $x_1(t)$ and $x_2(t)$, where $x_2(t)$ can be obtained as a filtered version of $x_1(t)$. In particular, we assume that

$$x_2(t) = (x_1 * h)(t)\,, \tag{1}$$

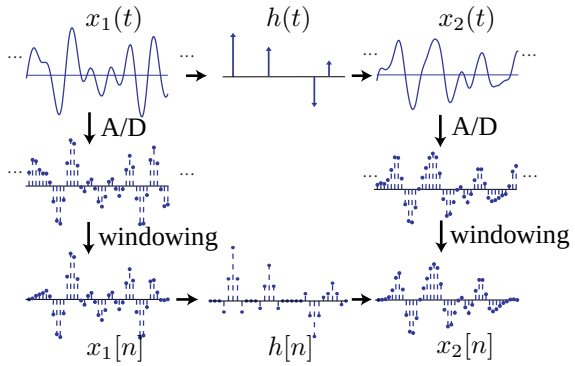


Figure 2: The continuous-time sparse filtering operation and its discrete-time counterpart.

where $h(t) = \sum_{k=1}^{K} c_k \delta(t - t_k)$ is a stream of K Diracs with *unknown* delays $\{t_k\}_{k=1}^K$ and coefficients $\{c_k\}_{k=1}^K$.
In this work, we study a finite-dimensional discrete version of the above model. As shown in Figure 2, we assume that the original continuous signal $x_1(t)$ is bandlimited to $[-\sigma, \sigma]$. Sampling $x_1(t)$ at uniform time interval T leads to a discrete sequence of samples $x_{s1}[n] \stackrel{\text{def}}{=} x_1(nT)$, where the sampling rate $1/T$ is set to be above the Nyquist rate σ/π. To obtain a finite-length signal, we subsequently apply a temporal window to the infinite sequence $x_{s1}[n]$ and get

$$x_1[n] \stackrel{\text{def}}{=} x_{s1}[n]\, w_N[n], \quad \text{for } n = 0, 1, ..., N-1,$$

where $w_N[n]$ is a smooth temporal window of length N. Note that when N is large enough, we can neglect the windowing effect, since $\widehat{w}_N(\omega)/(2\pi)$ approaches a Dirac function $\delta(\omega)$ as $N \to \infty$.
Applying the above procedure to $x_2(t)$ and using (1), we have

$$X_2[m] \approx \frac{1}{T}\, \widehat{x}_2\left(\frac{2\pi m}{NT}\right) \approx X_1[m] H[m], \qquad (2)$$

where

$$H[m] \stackrel{\text{def}}{=} \sum_{k=1}^{K} c_k e^{-j2\pi m t_k/(NT)}. \qquad (3)$$

The above relationship implies that the finite-length signals $x_1[n]$ and $x_2[n]$ can also be approximately modeled as the input and output of a *discrete-time* filtering operation[1]. In general, the location parameters $\{t_k\}$ in (3) can be arbitrary real numbers, and consequently, the discrete-time filter $h[n]$ is no longer sparse (see Figure 2 for a typical impulse response of $h[n]$). However, when the sampling interval T is small enough, we can assume that the real-valued delays $\{t_k\}$ are close enough to the sampling grid, i.e., $t_k/T \approx n_k$ for some integers $\{n_k\}$. We will follow this assumption[2] throughout the paper.

Definition 1 (Correlation Model) *The signals of interest are two vectors* $\boldsymbol{x}_1 = (x_1[0], \ldots, x_1[N-1])^T$ *and* $\boldsymbol{x}_2 = (x_2[0], \ldots, x_2[N-1])^T$, *linked to each other through a circular convolution*

$$x_2[n] = (x_1 \circledast h)[n] \quad \text{for } n = 0, 1, \ldots, N-1, \qquad (4)$$

where $\boldsymbol{h} = (h[0], \ldots, h[N-1])^T \in \mathbb{R}^N$ *is an unknown* K*-sparse vector, that is,* $\|\boldsymbol{h}\|_0 = K$.

3. Bounds

3.1 Universal Recovery

Let $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ be the sampling matrices used by the two sensors, and $\boldsymbol{A}$ be the block-diagonal matrix with $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ on the main diagonal. We first focus on finding those $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ such that every $\boldsymbol{x}^T = (\boldsymbol{x}_1^T, \boldsymbol{x}_2^T)$ is uniquely determined by its sampling data $\boldsymbol{Ax}$. Here we denote by $\mathcal{X}$ the set of all stacked vectors $\boldsymbol{x}$ such that its components $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ satisfy (4) for some K-sparse vector $\boldsymbol{h}$.

Definition 2 (Universal Achievability) *We say a sampling pair* (M_1, M_2) *is achievable for universal reconstruction if there exists fixed measurement matrices* $\boldsymbol{A}_1 \in \mathbb{R}^{M_1 \times N}$ *and* $\boldsymbol{A}_2 \in \mathbb{R}^{M_2 \times N}$ *such that the set*

$$\mathcal{B}(\boldsymbol{A}_1, \boldsymbol{A}_2) \stackrel{\text{def}}{=} \{\boldsymbol{x} \in \mathcal{X} : \exists\, \boldsymbol{x}' \in \mathcal{X} \text{ with } \boldsymbol{x} \neq \boldsymbol{x}' \text{ but } \boldsymbol{Ax} = \boldsymbol{Ax}'\} \qquad (5)$$

is empty.

Intuition suggests that, due to the correlation between the vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, the minimum number of samples needed to perfectly describe all possible vectors $\boldsymbol{x}$ can made smaller than the total number of coefficients $2N$. The following proposition shows that, surprisingly, this is not the case.

Proposition 1 *A sampling pair* (M_1, M_2) *is achievable for universal reconstruction if and only if* $M_1 \geq N$ *and* $M_2 \geq N$.

Proof Let us consider two stacked vectors $\boldsymbol{x}^T = (\boldsymbol{x}_1^T, \boldsymbol{x}_2^T)$ and $\boldsymbol{x}'^T = (\boldsymbol{x}_1'^T, \boldsymbol{x}_2'^T)$, each following the correlation model (4). They can be written under the form

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{I}_N \\ \boldsymbol{C} \end{bmatrix} \boldsymbol{x}_1 \quad \text{and} \quad \boldsymbol{x}' = \begin{bmatrix} \boldsymbol{I}_N \\ \boldsymbol{C}' \end{bmatrix} \boldsymbol{x}_1',$$

where $\boldsymbol{C}$ and $\boldsymbol{C}'$ are circulant matrices with vectors $\boldsymbol{h}$ and $\boldsymbol{h}'$ as the first column, respectively. It holds that

$$\boldsymbol{x} - \boldsymbol{x}' = \begin{bmatrix} \boldsymbol{I}_N & -\boldsymbol{I}_N \\ \boldsymbol{C} & -\boldsymbol{C}' \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_1' \end{bmatrix}.$$

Moreover, we have that

$$\text{rank} \begin{bmatrix} \boldsymbol{I}_N & -\boldsymbol{I}_N \\ \boldsymbol{C} & -\boldsymbol{C}' \end{bmatrix} = N + \text{rank}\,(\boldsymbol{C} - \boldsymbol{C}').$$

When $\boldsymbol{C} - \boldsymbol{C}'$ is of full rank, the above matrix is of rank $2N$. This happens, for example, when $K = 1$ with $\boldsymbol{C} = 2\boldsymbol{I}_N$ and $\boldsymbol{C}' = \boldsymbol{I}_N$. In this case, $\boldsymbol{x} - \boldsymbol{x}'$ can take any possible values in $\mathbb{R}^{2N}$. Hence, a necessary (and sufficient) condition for the set (5) to be empty is that the block-diagonal matrix $\boldsymbol{A}$ is a $M \times 2N$-dimensional matrix of full rank, with $M \geq 2N$. In particular, $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ must be full rank matrices of size $M_1 \times N$ and $M_2 \times N$, respectively, with $M_1, M_2 \geq N$. Note that, in the centralized scenario, the full rank condition would still require to take at least $2N$ measurements.

[1] Note that in order to be unambiguous in the positions $\{t_k\}$, we need to ensure that $NT > \max_k \{t_k\}$.

[2] We introduce this assumption (i.e. $t_k/T = n_k$ for some $n_k \in \mathbb{Z}$) mainly for the simplicity it brings to the theoretical analysis in later parts of this paper. It is however not an inherent limitation of our work.

3.2 Almost Sure Recovery

As shown in Proposition 1, universal recovery is a rather strong requirement to satisfy since we have to take at least N samples at each sensor, without being able to exploit the existing correlation. In many situations, however, it is sufficient to consider a weaker requirement, which aims at finding measurement matrices that permit the perfect recovery of *almost all* signals from $\mathcal{X}$.

Definition 3 (Almost Sure Achievability) *We say a sampling pair (M_1, M_2) is achievable for almost sure reconstruction if there exist* fixed *measurement matrices $\boldsymbol{A}_1 \in \mathbb{R}^{M_1 \times N}$ and $\boldsymbol{A}_2 \in \mathbb{R}^{M_2 \times N}$ such that the set $\mathcal{B}(\boldsymbol{A}_1, \boldsymbol{A}_2)$, as defined in* (5)*, is of probability zero.*

The above definition for the almost sure recovery depends on the probability distribution of the signal $\boldsymbol{x}_1$ and the sparse filter $\boldsymbol{h}$. In what follows, it is sufficient to assume that the signal $\boldsymbol{x}_1$ and the non-zero coefficients of the filter $\boldsymbol{h}$ have non-singular[3] probability distributions over $\mathbb{R}^N$ and $\mathbb{R}^K$, respectively. The following proposition gives an achievability bound of the number of samples needed for the almost sure reconstruction.

Proposition 2 *A sampling pair (M_1, M_2) is achievable for almost sure reconstruction if*

$$\begin{aligned} M_1 &\geq \min\{K + r, N\}, \\ M_2 &\geq \min\{K + r, N\}, \\ \textit{and} \quad M_1 + M_2 &\geq \min\{N + K + r, 2N\}, \end{aligned} \tag{6}$$

where $r = 1 + mod\,(K, 2)$.

Proof Due to space limitations, we just provide the sketch of the proof which is constructive in nature. First, let the two sensors take the Fourier transform of their signals and send the first $(K + r + 1)/2$ frequency components to the central decoder. By dividing the two sets of measurements (Note that the denominator should not be zero, which is guaranteed almost surely), the decoder calculates the necessary Fourier elements of the K-sparse filter $\boldsymbol{h}$ in order to reconstruct it almost surely. Then, the sensors transmit complementary subsets of frequency indices up to the Nyquist frequency. Knowing the filter $\boldsymbol{h}$ and the frequency content of one of the signals at some index, the decoder computes the corresponding frequency content of the other signal using (4).

Proposition 2 shows that, in contrast to the universal scenario, the correlation between the signals by means of the sparse filter provides a big saving in the almost sure setup, especially when $K \ll N$. This is depicted as the solid line in Figure 3.
Unfortunately, the algorithm that attains the bound in (6) is combinatorial in nature and thus, computationally prohibitive [1]. In the following, we propose a novel distributed sensing algorithm based on annihilating filters. This algorithm needs effectively K more measurements with respect to the achievability region for the almost sure reconstruction but exhibits polynomial complexity of $O(KN)$.

[3]By a non-singular distribution, we mean any continuous distribution such that the probability that the random variables lie in a low-dimensional subspace is zero.

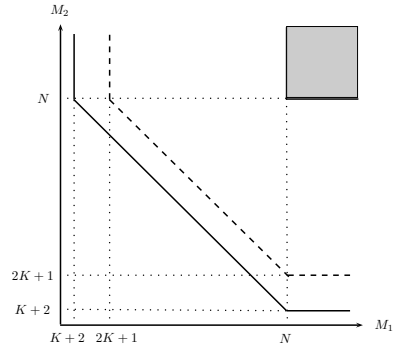


Figure 3: Achievable sampling region for universal reconstruction (shaded area), sampling pairs achieved for almost sure reconstruction for K odd (solid line) and sampling pairs achieved for almost sure reconstruction by the proposed algorithm based on annihilating filters (dashed line).

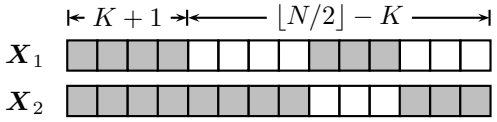


Figure 4: Sensors 1 and 2 both send the first $K + 1$ DFT coefficients of their observation, but only complementary subsets of the remaining frequency components.

4. Distributed Sensing Algorithm

The proposed distributed sensing scheme is based on a frequency-domain representation of the input signals. Let us denote by $\boldsymbol{X}_1 \in \mathbb{C}^N$ and $\boldsymbol{X}_2 \in \mathbb{C}^N$ the DFTs of the vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively. The circular convolution in (4) can be expressed as

$$\boldsymbol{X}_2 = \boldsymbol{H} \odot \boldsymbol{X}_1, \tag{7}$$

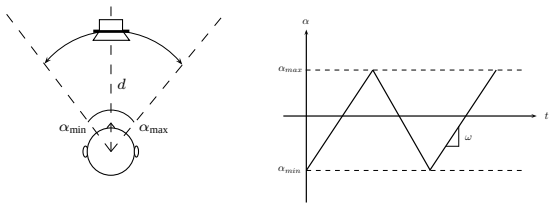
where $\boldsymbol{H} \in \mathbb{C}^N$ is the DFT of the filter $\boldsymbol{h}$ and $\odot$ denotes the element-wise product. Our approach consists of two main steps:

1. Finding filter $\boldsymbol{h}$ by sending the first $K + 1$ (1 real and K complex) DFT coefficients of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.
2. Sending the remaining frequency indices by sharing them among the two sensors.

The decoder first finds the filter $\boldsymbol{h}$ using only the first $K+1$ DFT coefficients of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. To this end, the decoder first computes

$$H[m] = \frac{X_2[m]}{X_1[m]} \quad \text{and} \quad H[-m] = H^*[m] \tag{8}$$

provided that $X_1[m]$ is non-zero for $m = 0, 1, \ldots, K$. This happens almost surely if the distribution of $\boldsymbol{x}_1$ is, for example, non-singular. Then, it finds the K-sparse filter with an annihilating filter approach; see [7] for details. The senors also transmit complementary subsets (in terms of frequency indexes) of the remaining DFT coefficients of their signals ($N - 2K - 1$ real values in total). This is illustrated in Figure 4. The first $K + 1$ DFT coefficients allow to almost surely reconstruct the filter $\boldsymbol{h}$. The missing frequency components of $\boldsymbol{x}_1$ (resp. $\boldsymbol{x}_2$) are then recovered from the available DFT coefficients of $\boldsymbol{x}_2$ (resp. $\boldsymbol{x}_1$) using the relation (7).



(a) (b)

Figure 5: Audio Experiment Setup. (a) A sound source travels at a distance of d meter in front of the head. (b) Angular position of the sound source with respect to time.

Note that in order to compute $X_1[m]$ from $X_2[m]$, the frequency components of the filter $H[m]$ should be nonzero. This is insured almost surely with our assumption that the nonzero elements of the filter $\boldsymbol{h}$ are chosen according to a non-singular distribution in $\mathbb{R}^K$. In terms of achievability, we have thus shown the following result.

Proposition 3 *A sampling pair (M_1, M_2) is achievable for almost sure reconstruction using the efficient annihilating filter method if*

$$\begin{aligned} M_1 &\geq \min\{2K+1, N\}, \\ M_2 &\geq \min\{2K+1, N\}, \\ \textit{and} \quad M_1 + M_2 &\geq \min\{N+2K+1, 2N\}. \end{aligned}$$

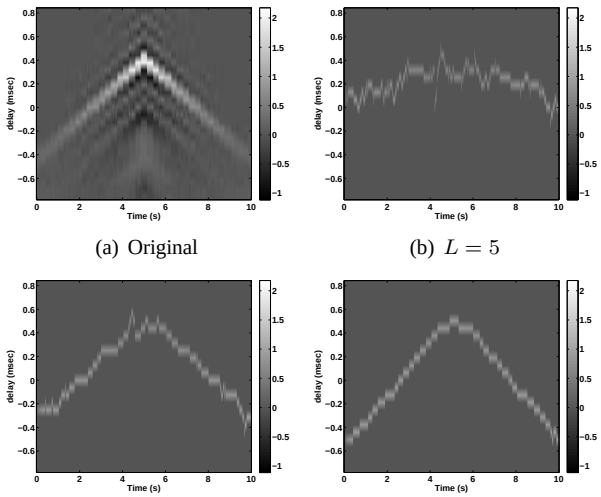
In the presence of noise or model mismatch, we add robustness to the system by sending $L+1$ DFT coefficients of x_i $(i = 1, 2)$ with $L \geq K$ to the decoder. We denoise the measurements by using the denoising algorithm due to Cadzow; for details see [3]. Then the annihilating filter method uses the denoised measurements to estimate the sparse filter.

5. Application

In a practical scenario, we consider the signals recorded by two hearing aids mounted on the left and right ears of the user. We assume that the signals of the two hearing aids are related thorough a filtering operation. We refer to this filter as binaural filter. In the presence of a single source in far field, and neglecting reverberations and the head-shadow effect [2], the signal recorded at hearing aid 2 is simply a delayed version of the one observed at hearing aid 1. Hence, the binaural filter can be assumed to have sparsity factor $K = 1$. In the presence of reverberations and head shadowing, the filter from one microphone to the other is no longer sparse which introduces model mismatch. Despite this model mismatch, the transfer function between the two received signals should be approximately sparse, with the main peak indicating the desired relative delay.

In our setup, a single sound source located at distance $d = 1$ meter from the head of a KEMAR mannequin, moves back and forth between two angles $\alpha_{\min} = -45^\circ$ and $\alpha_{\max} = 45^\circ$. The angular speed of the source is $\omega = 18$ deg/sec. The sound is recorded by the microphones of the two hearing aids, located at the ears of the mannequin. We want to retrieve the binaural filter between the two hearing aids at hearing aid 1, from limited data transmitted by hearing aid 2. Then, the main peak of the binaural filter indicates the relative delay between the two received signals, which can be used to localize the source.

Figure 6 demonstrates the localization performance of the algorithm. Figure 6(a) shows the evolution of the original binaural impulse response over time. Figures 6(b)- 6(d) exhibits the sparse approximation to the filter, using different number of measurements. This clearly demonstrates the effect of the over-sampling factor on the robustness of the reconstruction algorithm.



(a) Original (b) $L = 5$

(c) $L = 15$ (d) $L = 25$

Figure 6: Tracking the binaural impulse response. Each column in the image corresponds to the binaural impulse response at the time mentioned on the x axis. (a) Original binaural filter. (b)-(d) Tracking the evolution of the main peak with different values of the oversampling factor L.

6. Conclusions

A general formulation of the distributed sensing problem has been proposed where the two signals are connected through an unknown sparse filter. In this context, both universal and almost sure reconstruction were addressed together with their corresponding achievable bounds. In addition, a distributed sensing scheme was presented, together with a method to make it robust to model mismatch. Our future research will focus on investigating more the applications of the proposed methods in the distributed sensing context.

References:

[1] D. Baron, M. B. Wakin, M. F. Duarte, S. Sarvotham, and R. G. Baraniuk. Distributed compressed sensing. Technical Report ECE-0612, Electrical and Computer Engineering Department, Rice University, Dec. 2006.

[2] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization.* MIT Press, Cambridge, MA, 1997.

[3] J. A. Cadzow. Signal enhancement – A composite property mapping algorithm. *IEEE Trans. Acoust., Speech, Signal Process.*, 36(1):49–67, Jan. 1988.

[4] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2):489–509, Feb. 2006.

[5] D. L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306, Apr. 2006.

[6] D. Slepian and J. K. Wolf. Noiseless coding of correlated information sources. *IEEE Trans. Inf. Theory*, 19:471–480, Jul. 1973.

[7] M. Vetterli, P. Marziliano, and T. Blu. Sampling signals with finite rate of innovation. *IEEE Trans. Signal Process.*, 50(6):1417–1428, Jun. 2002.

A method for generalized sampling and reconstruction of finite-rate-of-innovation signals

Chandra Sekhar Seelamantula and Michael Unser
Biomedical Imaging Group
Ecole polytechnique fédérale de Lausanne
Switzerland
{chandrasekhar.seelamantula, michael.unser}@epfl.ch

Abstract:

We address the problem of generalized sampling and reconstruction of finite-rate-of-innovation signals. Specifically, we consider the problem of sampling streams of Dirac impulses and propose a two-channel method that enables fast, local reconstruction under some suitable conditions. We also specify some acquisition kernels and give the associated reconstruction formulas. It turns out that these kernels can also be combined into one kernel, which can be employed in the single-channel sampling scenario. The single-kernel approach carries over all the advantages of the two-channel counterpart. Simulation results are presented to validate the theoretical calculations.

1. Introduction and prior art

Sampling theory is the foundation on which digital signal processing has been built. The popular flavor of the sampling theory is due to Shannon [1] and deals exclusively with bandlimited signals. Shannon's theory was generalized in several ways, the most prominent one being the theory of multichannel sampling developed by Papoulis [3]—his theory is known as the Generalized Sampling Theory. Papoulis' formalism, however, deals only with bandlimited signals. To accommodate the more general class of finite-energy signals, Unser and Zerubia [4] developed a theory, which does not rely on the bandlimiting constraint. Another important extension is the sampling and reconstruction of signals that lie in some shift-invariant subspace spanned by the integer-shifted versions of a generator kernel (see [2] and the references therein). The specific case of bandlimited sampling corresponds to a sinc kernel and is subsumed by this formalism.

Recently, Vetterli *et al.* [5] extended sampling theory in a new direction to answer a question that has not been addressed before—that of sampling and reconstructing streams of Dirac impulses and signals derived therefrom. These signals are not constrained to lie in the space of finite-energy functions nor in the space of bandlimited functions. They may also not lie in some shift-invariant subspace generated by a kernel. Typically, such signals are specified by a set of discrete parameters per time unit, also known as their rate of innovation. We are interested in signals that have a finite rate of innovation (FRI). Specifically, consider the stream of time-ordered Dirac impulses:

$$x(t) = \sum_{\ell=1}^{L} a_\ell \delta_D(t - t_\ell), \tag{1}$$

where $\delta_D(\cdot)$ denotes the Dirac impulse. The problem is to compute the parameters $\{a_\ell, t_\ell; 1 \leq \ell \leq L\}$ based on some measurements on $x(t)$. The parametric nature of the problem has resulted in the development of techniques that are quite different from those that sampling theorists have been familiar with. Typically, the reconstruction techniques developed by Vetterli *et al.* [5] and Dragotti *et al.* [6] have a flavor of parametric spectral estimation [7]. They also employ in a novel fashion spline kernels [8, 9] that reproduce polynomials or exponentials. It is remarkable that these kernels, which play a vital role in wavelet theory, are also quite useful for sampling FRI signals.

In the techniques mentioned above, the focus is exclusively on the single-channel case. Recently, some new multichannel approaches have also been developed. Kusuma and Goyal proposed a new technique for reconstructing an unknown number of impulses over a finite interval of time by using a successive approximation criterion [10]. Their technique can be implemented using a bank of integrators and B-splines. Baboulaz and Dragotti proposed a distributed acquisition scheme for FRI signals and demonstrated applications to image registration and super-resolution image restoration [11]. In [12], we have proposed a two-channel sampling method for the FRI problem (cf. Fig. 1). We have employed first-order resistor-capacitor networks to sample streams of Dirac impulses and piecewise-constant functions. The reconstruction technique boils down to solving a system of two equations containing the unknown parameters in decoupled form. The key result in [12] is given below:

Proposition 1 *The stream of Dirac impulses in* (1) *is uniquely specified by the samples* $y_\alpha(nT) = (x * h_\alpha)(nT)$ *and* $y_\gamma(nT) = (x * h_\gamma)(nT)$, $n \in \mathbb{Z}$, *where* $h_\alpha(t) = \alpha\, e^{-\alpha t} u(t)$, $h_\gamma(t) = \gamma\, e^{-\gamma t} u(t)$, *and* $\alpha \neq \gamma$, *provided that* $\min_{2 \leq \ell \leq L} \{t_\ell - t_{\ell-1}\} \geq T$.

1.1 Motivation for the present work

The above proposition relies on causal exponential functions for sampling. Working with exponentials has the practical advantage that they can be easily generated by employing first-order resistor-capacitor circuits. From a mathematical viewpoint, however, exponentials are probably not the only class of functions that enable accurate reconstruction. The main motivation behind the present paper is the quest for alternative kernels $h_\alpha(t)$ and $h_\gamma(t)$ that would fit into the framework of the above proposition (also cf. Fig. 1). To that end, we first reformulate the method proposed in [12] in a more general framework and specify some kernels that enable exact reconstruction.

2. Generalized sampling formulation

Consider the two-channel sampling scenario shown in Fig. 1. Let $h_\alpha(t)$ and $h_\gamma(t)$, $\alpha, \gamma \in \mathbb{C}$, denote the impulse responses of two causal linear shift-invariant systems, compactly supported on $[0, T]$ and nonzero over that interval. Consider the stream of Dirac impulses in (1), where the impulses are separated by at least T; i.e.,

$$\min_{2\leq \ell \leq L} \{t_\ell - t_{\ell-1}\} \geq T. \tag{2}$$

Deviations from this condition shall be addressed later. The output of the system to the input $x(t)$ is given by

$$y_\alpha(t) \triangleq (x * h_\alpha)(t) = \sum_{\ell=1}^{L} a_\ell h_\alpha(t - t_\ell).$$

Let us next consider the samples of $y_\alpha(t)$ taken on a uniform grid with a sampling step T. Note that we have chosen the sampling period to be equal to the support of $h_\alpha(t)$; otherwise, we are likely to miss some closely-spaced impulses as the following analysis shows. The samples of $y_\alpha(t)$ are given by

$$y_\alpha(nT) = \sum_{\ell=1}^{L} a_\ell \, h_\alpha \left(nT - t_\ell\right) \delta_K \left[nT - r(t_\ell)\right],$$

where $r(t_\ell) = \left\lceil \frac{t_\ell}{T} \right\rceil T$ is the operator that performs the ceiling of t_ℓ with respect to the sampling grid and δ_K denotes the Kronecker impulse. The sequence $y_\alpha(nT)$ comprises Kronecker impulses, each corresponding to a Dirac impulse in $x(t)$ under the condition (2). Note that the sampling period T equals the support of the kernel. Similarly, corresponding to a system with impulse response $h_\gamma(t)$, $\gamma \neq \alpha$, we have

$$y_\gamma(nT) = \sum_{\ell=1}^{L} a_\ell \, h_\gamma \left(nT - t_\ell\right) \delta_K \left[nT - r(t_\ell)\right].$$

Note that these sampling instants correspond to the nonzero values in the sequences $y_\alpha(nT)$ and $y_\gamma(nT)$ and are therefore known. Consider the ℓ^{th} nonzero samples in the sequences $y_\alpha(nT)$ and $y_\gamma(nT)$:

$$y_\alpha(r(t_\ell)) = a_\ell \, h_\alpha[r(t_\ell) - t_\ell] \text{ and} \tag{3}$$
$$y_\gamma(r(t_\ell)) = a_\ell \, h_\gamma[r(t_\ell) - t_\ell]. \tag{4}$$

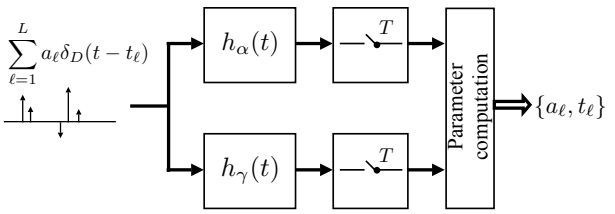


Figure 1: Two-channel sampling of a stream of dirac impulses.

In (3) and (4), the indices $r(t_\ell)$ and the values on the left hand side are known. The impulse responses h_α and h_γ are also known; their design shall be explained below. The amplitude and position parameters $\{t_\ell, a_\ell\}$ are unknown and have to be determined. The amplitude of the ℓ^{th} Dirac impulse appears as a multiplicative factor. The position of the Dirac impulse is encoded in the amplitude of the Kronecker impulse. Dividing (3) by (4) eliminates a_ℓ and gives rise to an equation in the unknown t_ℓ, which can be computed if and only if $(h_\alpha / h_\gamma)(t)$ is invertible on its range. The value of t_ℓ thus obtained can then be substituted in (3) or (4) to obtain the value of a_ℓ. Some specific functions that fit into the above reconstruction paradigm are presented next.

3. Kernels for two-channel sampling

We specify only the kernel $h_\alpha(t)$; unless otherwise mentioned, $h_\gamma(t)$ is obtained by replacing α with γ; i.e., both kernels have the same functional form. The kernels involve gating by the B-spline of order zero, at scale T: $\beta(t) = u(t) - u(t - T)$, where $u(t)$ is the unit step function. We specify the kernel definitions and give the expressions for $\{t_\ell, a_\ell\}$ directly. The intermediate calculations are omitted but it is straightforward to supply them starting from the definition of the kernel.

1. Exponential spline (E-spline) kernels [9]: $h_\alpha(t) = e^{-\alpha t}\beta(t), \alpha \in \mathbb{R}$, where $u(t)$ is the unit-step function. The parameters of ℓ^{th} impulse are given by

$$t_\ell = r(t_\ell) + \frac{1}{\alpha - \gamma} \log\left(\frac{y_\alpha(r(t_\ell))}{y_\gamma(r(t_\ell))}\right) \text{ and}$$
$$a_\ell = y_\alpha(r(t_\ell)) \exp\left(-\frac{\alpha}{\alpha - \gamma} \log\left(\frac{y_\alpha(r(t_\ell))}{y_\gamma(r(t_\ell))}\right)\right).$$

   This kernel choice has been analyzed in sufficient detail in [12]. The specific kernel given above is a first-order E-spline kernel. One could, in principle, also employ higher-order kernels. The advantage of first-order E-spline kernels over the higher-order ones, however, is that they always give rise to closed-form solutions. The higher-order kernels exhibit this property only for certain values of the spline parameters. For further discussion on this issue, we refer the reader to [12].

2. Power functions: $h_\alpha(t) = t^\alpha \beta(t)$, $\alpha \in \mathbb{R}$. Corre-

spondingly, the parameters of $x(t)$ are given by

$$t_\ell = r(t_\ell) - \left(\frac{y_\alpha(r(t_\ell))}{y_\gamma(r(t_\ell))}\right)^{\frac{1}{\alpha-\gamma}} \text{ and}$$
$$a_\ell = y_\alpha(r(t_\ell)) \left(\frac{y_\alpha(r(t_\ell))}{y_\gamma(r(t_\ell))}\right)^{\frac{-\alpha}{\alpha-\gamma}}.$$

For $\alpha \in \mathbb{Z}^+$, the power function becomes a monomial of degree α. Since B-splines of order α can reproduce polynomials (and naturally, monomials too) up to degree α, they are included as special elements of this class. Therefore, power functions, which play a vital role in moment-based sampling approaches [6, 11] for the FRI problem, are also useful in the generalized sampling approach. Also, note that fractional powers are admissible in the kernel definition.

3. Gaussian functions: $h_\alpha(t) = e^{-\alpha t^2}\beta(t)$, where $\alpha \in \mathbb{R}$. Correspondingly, we have that

$$t_\ell = r(t_\ell) - \sqrt{\frac{1}{\alpha-\gamma}\log\left(\frac{y_\gamma(r(t_\ell))}{y_\alpha(r(t_\ell))}\right)}, \text{and}$$
$$a_\ell = \exp\left(\frac{\alpha}{\alpha-\gamma}\log\left(\frac{y_\gamma(r(t_\ell))}{y_\alpha(r(t_\ell))}\right)\right).$$

4. Complex E-splines: $h_\alpha(t) = e^{-\mathrm{j}\alpha t}\beta(t)$, $\alpha \in \mathbb{R}$. This kernel cannot be treated as a special case of the E-spline kernels with an imaginary parameter. The reason is that there is an issue related to *parameter identifiability* that deserves special attention. The potential problem is that this kernel may give rise to more than one solution for t_ℓ; there is, however, no ambiguity in the solution for a_ℓ. We further explain this issue and also state a condition that helps overcome the non-uniqueness hurdle.
The cause of ambiguity is essentially the quasi-periodicity of the complex exponential over the support $[0, T]$:

$$e^{-\mathrm{j}\alpha\,(r(t_\ell)-t_\ell)} = e^{-\mathrm{j}\alpha\left(r(t_\ell)-t_\ell+\frac{2m\pi}{\alpha}\right)},$$

for $m \in \mathbb{Z}$ such that $0 \leq (r(t_\ell)-t_\ell+\frac{2m\pi}{\alpha}) \leq T$. The restriction on m is due to the fact that we are considering a truncated complex exponential. The inequality gives rise to multiple solutions for t_ℓ. The solution to this problem lies in tying up the choices of the values of α and T such that $m = 0$ is the only possibility in the above inequality. This amounts to requiring that the complex exponential have at maximum one period within a sampling interval; i.e., $\frac{2\pi}{\alpha} > T$. Under this condition, we have the reconstruction formulae:

$$t_\ell = -\mathrm{j}\,\log\left(\frac{y_\alpha(r(t_\ell))e^{\mathrm{j}\alpha\,r(t_\ell)}}{y_\gamma(r(t_\ell))e^{\mathrm{j}\gamma\,r(t_\ell)}}\right), \text{and}$$
$$a_\ell = y_\alpha(r(t_\ell))\exp\left(\mathrm{j}\alpha(r(t_\ell)-t_\ell)\right).$$

Similarly, a truncated Fresnel kernel can be employed by considering purely imaginary parameters in the definition of the Gaussian above. For complex parameters, the E-spline and Fresnel kernels have an exponential and Gaussian decay, respectively.

5. Hybrid sampling kernels: In the kernels considered above, we have enforced the same functional form for both $h_\alpha(t)$ and $h_\gamma(t)$. By relaxing this property, we can make the reconstruction technique more efficient. For example, if we set one of the parameters (but not both), say α to zero, the kernel reduces to a causal B-spline of order 0: $h_\alpha(t) = \beta(t)$. The second kernel can be taken from any of the choices listed above. The samples from the zeroth-order B-spline channel then directly yield $a_\ell = y_\alpha(r(t_\ell))$. Using the samples from the second channel, we can compute the positions of the Dirac impulses. For example, if we employ the truncated power function in the second channel, we have that $t_\ell = r(t_\ell) - \left(\frac{y_\gamma(r(t_\ell))}{a_\ell}\right)^{\frac{1}{\gamma}}$. Note that $r(t_\ell)$ and $y_\gamma(r(t_\ell))$ are known.

Having listed a few kernel choices, we reiterate that, in the present formalism, the condition stated in (2) is crucial for the super-resolution localization of impulses. If two successive Dirac impulses are spaced closer apart than the sampling period, then they give rise to overlapping Kronecker impulses and resolving them is not possible within the proposed formulation. The existing approaches [5, 6, 10, 11] do not suffer from this limitation.

4. Kernels for single-channel sampling

The principal advantage offered by the two-channel method equipped with the choice of a proper kernel is the decoupling between the amplitudes and positions of the impulses. As shown next, this advantage can be carried over to the single-channel case by suitably integrating the previously listed kernels into a single function. For example, consider the kernel: $h_{\alpha,\gamma}(t) = e^{-\alpha t}\beta(t) + e^{-\gamma(t-T)}\beta(t-T)$, which has the same properties as the hybrid kernel in the two-channel case (kernel (1) in Section 3.). This choice would give rise to two nonzero samples per Dirac impulse, which can be used to solve for a_ℓ and t_ℓ. Again, if $\alpha = 0$, the first sample would straightaway give the amplitude, which can then be used together with the second sample to compute the position. Thus, we have a similar algorithm as in the two-channel case, the only difference being that, in the two-channel case, these samples are acquired one per channel whereas in the one-channel case, they are acquired in the same channel—the overall sampling rate, however, is the same in both the cases. In general, the kernels for the single-channel case can be defined as: $h_{\alpha,\gamma}(t) = h_\alpha(t) + h_\gamma(t-T)$. Since the support of the kernel $h_{\alpha,\gamma}(t)$ is double that of $h_\alpha(t)$ or $h_\gamma(t)$, impulses that are farther apart by at least $2T$ $\left(\text{i.e., } \min_{2\leq\ell\leq L}\{t_\ell - t_{\ell-1}\} \geq 2T\right)$ only can be resolved. The kernels defined in this paper are shown in Fig. 2.

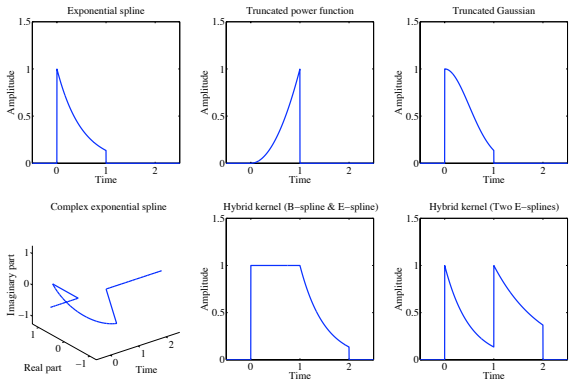


Figure 2: Sampling kernels. The parameters $\alpha = 2, \gamma = 1$, and $T = 1$, are chosen for the sake of illustration.

5. Simulations

We next validate the theoretical findings by numerical experiments. We simulate the two-channel sampling of nine Dirac impulses shown in Fig. 3(a); the amplitudes and positions are chosen for the purpose of illustration. The minimum spacing between two impulses is 0.0076 seconds. The sampling period T is chosen to be 0.0038 seconds to ensure that (2) is satisfied. The impulses are sampled using the power function kernels with parameters $\alpha = 3, \gamma = 2$, and $T = 0.0038$ seconds. These values are chosen for the purpose of illustration. The reconstructed stream of Dirac impulses is shown in Fig. 3(b). The reconstruction is accurate to numerical precision. Identical results were obtained with the other kernel choices.

6. Conclusions

We have extended the results developed in [12] and proposed new kernels for both single-channel and two-channel sampling scenarios. The kernels are built using functions known in system theory such as the exponential, power function, Gaussian, etc. The main advantage of the proposed formulation is that, under the condition of minimum separation between consecutive impulses, a fast local reconstruction algorithm can be developed. This advantage, however, comes with the shortcoming that impulses spaced farther apart than the sampling period only can be resolved. It would be a challenging task to develop local reconstruction algorithms without imposing constraints on the minimum/average separation between impulses or groups thereof.

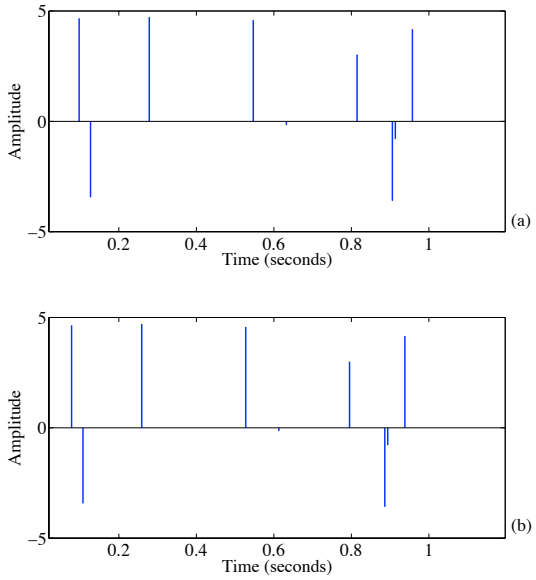


Figure 3: (a) Ground truth, (b) Reconstructed signal.

Acknowledgments

This work was supported by the Swiss National Science Foundation (SNSF) Grant 200020-101821.

References:

[1] C. E. Shannon, "Communication in the presence of noise," *Proc. IRE*, vol. 37, no. 1, pp. 10-21, Jan. 1949.

[2] M. Unser, "Sampling—50 years after Shannon," *Proc. IEEE*, vol. 88, no. 4, pp. 569-587, Apr. 2000.

[3] A. Papoulis, "Generalized sampling expansion," *IEEE Trans. Circuits Syst.*, vol. 24, no. 11, pp. 652–654, 1977.

[4] M. Unser and J. Zerubia, "A generalized sampling theory without band-limiting constraints," *IEEE Trans. Circuits Syst. II, Analog and Digit. Signal Process.*, vol. 45, no. 8, pp. 959–969, Aug. 1998.

[5] M. Vetterli, P. Marziliano, and T. Blu, "Sampling signals with finite rate of innovation," *IEEE Trans. Signal Process.*, vol. 50, no. 6, pp. 1417–1428, Jun. 2002.

[6] P.L. Dragotti, M. Vetterli, and T. Blu, "Sampling moments and reconstructing signals of finite rate of innovation: Shannon meets Strang-Fix," *IEEE Trans. Signal Process.*, vol. 55, no. 5, pp. 1741–1757, May 2007, Part 1.

[7] P. Stoica and R. Moses, *Introduction to Spectral Analysis*, Englewood Cliffs, NJ: Prentice-Hall, 2000.

[8] M. Unser, "Splines: A perfect fit for signal and image processing," *IEEE Signal Process. Mag.*, vol. 16, no. 6, pp. 22–38, Nov. 1999.

[9] M. Unser and T. Blu, "Cardinal exponential splines: Part I—Theory and filtering algorithms," *IEEE Trans. Signal Process.*, vol. 53, no. 4, pp. 1425–1438, Apr. 2005.

[10] J. Kusuma and V. K. Goyal, "Multichannel sampling of parametric signals with a successive approximation property," in *Proc. IEEE Intl. Conf. on Imag. Proc.*, 2006, pp. 1265–1268.

[11] L. Baboulaz and P. L. Dragotti, "Distributed acquisition and image super-resolution based on continuous moments from samples," in *Proc. IEEE Intl. Conf. on Imag. Proc.*, 2006, pp. 3309–3312.

[12] C. S. Seelamantula and M. Unser, "A generalized sampling method for finite-rate-of-innovation-signal reconstruction," *IEEE Signal Process. Lett.*, vol. 15, pp. 813-816, 2008.

MULTICHANNEL SAMPLING OF TRANSLATED, ROTATED AND SCALED BILEVEL POLYGONS USING EXPONENTIAL SPLINES

Hojjat Akhondi Asl and Pier Luigi Dragotti

Imperial College London
Department of Electrical and Electronic Engineering
hojjat.akhondi-asl@imperial.ac.uk, p.dragotti@imperial.ac.uk

ABSTRACT

Recently there has been an interest in single and multichannel sampling of certain parametric signals based on rate of innovation using exponential reproducing kernels. In [5] it was shown that, using exponential reproducing kernels, we can achieve a fully symmetric multichannel sampling system where different channels receive translated versions of the input signal. For the case of bilevel polygons as the input signal considered in [5], having only translations is not practical and one may want to look at the cases of more complicated geometric transformations, such as rotation and scaling. In this paper we present a sampling theorem for multichannel sampling of translated, rotated and scaled bilevel polygons using Radon projections and generalized exponential splines.

1. INTRODUCTION

Recently, it was shown [1, 2] that it is possible to sample and perfectly reconstruct some classes of non-bandlimited signals using suitable sampling kernels. Signals that can be reconstructed using this framework are called signals with Finite Rate of Innovation (FRI) as they can be completely defined by a finite number of parameters. Stream of weighted Dirac impulses and bilevel polygons are some examples of FRI signals.

There has been a recent interest in sampling FRI signals using exponential spline [3] (E-spline) kernels. Dragotti et al. [2] showed that E-splines can be used as the sampling kernel to sample and perfectly reconstruct 1-D FRI signals. Extensions to the multidimensional case were considered in [5, 14] where we proposed sampling theorems for a stream of 2-D Dirac impulses (based on the ACMP algorithm [11]) and bilevel polygons (based on Radon projections [10]). Apart from the sampling kernels used in [5, 14], the reconstruction algorithms are also different from the ones used in the conventional multidimensional sampling theories [12, 13].

An advantage of E-spline sampling kernels over polynomial reproducing kernels such as B-splines is that, they can be employed in a fully symmetric multichannel sampling environment. By symmetric sampling, we mean that the sampling process can be evenly distributed between different acquisition devices. The inspiration and development of multichannel sampling of FRI signals is very recent and it has been looked at in [5, 6, 7, 8].

In [6] Seelamantula and Unser, by using simple RC filters, propose a simple acquisition and reconstruction method within the framework of multichannel sampling, where 1-D FRI signals such as an infinite stream of nonuniformly-spaced Dirac impulses and piecewise-constant signals can be sampled and perfectly reconstructed. In [7] Kusuma and Goyal proposed new ways of sampling 1-D Dirac impulses using a bank of integrators or B-splines. Their proposed scheme is closely related to previously known cases [1, 2] but provides a successive approximation property, which could be useful for detecting undermodelling when the number of Dirac impulses are unknown. In [8] Baboulaz and Dragotti use a multichannel sampling setup for sampling FRI signals and utilize that for image registration based on continuous moments and image super-resolution.

In [5] we illustrate that symmetric multichannel sampling of bilevel polygons can be achieved with the geometric transformations being a 2-D translation between the different signals. In practice, this is usually not the case, and in this paper we want to look at the cases of more complicated geometric transformations, such as rotation and scaling. The paper is organised as follows: In Section II we will briefly discuss the sampling setup needed for sampling 2-D FRI signals (single channel) and based on that we describe our multichannel sampling setup. In Section III we present our algorithm for sampling and perfectly reconstructing translated, rotated and scaled bilevel polygons with the use of generalized E-splines and Radon projections. In Section IV we provide simulation results to support our proposed theory.

2. MULTICHANNEL SAMPLING SETUP

Before describing the multichannel sampling framework, let us first, for the sake of clarity, show how a general 2-D sampling setup (single channel) for FRI signals is represented. Figure 1 shows a general 2-D sampling setup for FRI signals

where $g(x,y)$ represents the input signal, $\varphi(x,y)$ the sampling kernel, $s_{j,k}$ the samples and Tx, Ty are the sampling intervals. From the setup shown in Figure 1, the samples $s_{j,k}$

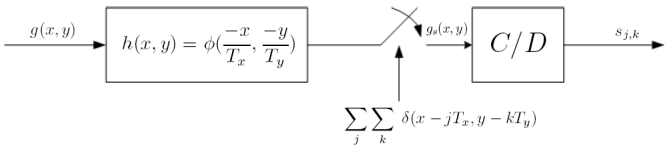


Fig. 1. 2-D sampling setup

are given by:

$$s_{j,k} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)\ \varphi(\frac{x}{T_x} - j, \frac{y}{T_y} - k)\ dx\ dy \quad (1)$$

where the kernel $\varphi(x,y)$ is the time reversed version of the filter response $h(x,y)$. $\varphi(x,y)$ can easily be produced by the tensor product between $\varphi(x)$ and $\varphi(y)$, that is $\varphi(x,y) = \varphi(x) \otimes \varphi(y)$. As mentioned before, $\varphi(x,y)$ is chosen to be an exponential reproducing kernel. The theory of exponential reproducing kernels is quite recent and is based on the notion of exponential splines (E-splines) [3]. A function $\beta_{\vec{\alpha}}(x)$ with Fourier transform

$$\hat{\beta}_{\vec{\alpha}}(\omega) = \prod_{n=0}^{N} \frac{1 - e^{\alpha_n - j\omega}}{j\omega - \alpha_n}$$

is called E-spline of order N where $\vec{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_N)$ can be real or complex. The produced spline has a compact support and can reproduce any exponential in the subspace spanned by $(e^{\alpha_0 t}, e^{\alpha_1 t}, \ldots, e^{\alpha_N t})$ which is obtained by successive convolutions of lower order E-splines ((N+1)-fold convolution). Exponential spline kernels can therefore reproduce, with their shifted versions, real or complex exponentials. That is, in 2-D form, any kernel satisfying:

$$\sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} c_{j,k}^{m,n} \varphi(x - j, y - k) = e^{\alpha_m x} e^{\beta_n y} \quad (2)$$

is an E-spline for a proper choice of the coefficients $c_{j,k}^{m,n}$. Here, $m = 0, 1, \ldots, M$, $n = 0, 1, \ldots, N$, $\alpha_m = \alpha_0 + m\lambda_1$ and $\beta_n = \beta_0 + n\lambda_2$. The values of (α_0, β_0) and (λ_1, λ_2) can be chosen arbitrarily, but too small or too large values could lead to unstable results for the reproduction of exponentials. E-splines are biorthogonal functions and the coefficients $c_{j,k}^{m,n}$ can be found using the dual of $\beta_{\vec{\alpha}}(x)$. An important property of E-splines is that they are a generalized version of B-splines. This is because, if the $\vec{\alpha}$ parameters are set to zero, then the produced spline would result in a B-spline, a polynomial reproducing spline. This property will be used to estimate the transformation parameters in Section III. The reader can refer to [5, 14] for sampling theories on single-channel sampling and perfect reconstruction of 2-D Dirac impulses and bilevel polygons using exponential splines.

We can now describe our multichannel sampling setup. A multichannel sampling system can be thought of multiple acquisition devices observing an input signal. In order to perfectly reconstruct the input signal using only one acquisition device, we normally require expensive acquisition devices with high sampling rates. By using a bank of acquisition devices (filters) and synchronizing the different channels exactly, we are able to reduce the number of samples needed from each device, resulting in a cheaper and more efficient sampling system. To model our multichannel system, consider a bank of E-spline filters to acquire FRI signals where each filter has access to a geometrically transformed version of the input signal. Figure 2 shows the described multichannel sampling scenario where the bank of filters $\varphi_1(x,y), \varphi_2(x,y), \ldots, \varphi_{N-1}(x,y)$ receive different versions of the input signal $g_0(x,y)$. Here, the geometric transformations (e.g. translation, rotation and scaling) are denoted by $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_{N-1}$.

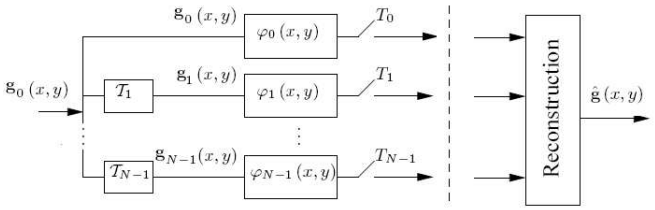


Fig. 2. Multichannel sampling setup

In [4] Baboulaz considered the use of E-splines for sampling a stream of 1-D Dirac impulses in a multichannel sampling setup described in Figure 2. He showed that if two 1-D signals are just shifted version of the other, then by setting one parameter to be common between the exponents of the E-spline sampling kernels for the two signals, one can not only estimate the shifts between the two signals, but also can linearly relate the exponential moments of the two signals (the reader can refer to [4, 5, 14] for more detailed discussion). Because of the direct relationship between the exponential moments of the two signals, we can achieve perfect reconstruction of the reference signal with fewer exponential moments required. Since less moments are required from each signal, a lower order E-spline sampling kernel would be needed, which in turn less samples from each signal are required to achieve perfect reconstruction. This is because, from [2] we know that a stream of Dirac impulses is uniquely determined from the samples if there are at most K Dirac impulses in an interval size of $2KLT$ where L is the support of the sampling kernel. Since the support of the sampling kernels is reduced in the multichannel case, we can achieve the same performance with a smaller sampling rate T.

3. ALGORITHM

Unfortunately we can not estimate the more complicated geometric transformations like the way it was done for the simple translation case in [5] with exponential reproducing kernels. Also, even if we assume that the transformation parameters are known and given, we still can not use the sampling algorithm shown in [5] for the multichannel framework. This is because introducing more complicated transforms such as rotation and/or scaling for example, would result in a non-linear relationship between the exponential moments of the different signals.

The first question we need to answer is that, assuming an oracle gives us the values of the transformation parameters, can we sample and perfectly reconstruct translated, rotated and scaled bilevel polygons in a symmetric multichannel framework? It is known that for an N-sided bilevel polygon, with N+1 projections, perfect reconstruction of the polygon can be achieved. That is points that have N+1 line intersections from the N+1 back-projections correspond to the N vertices of the polygon [9]. We also know that a Radon projection at an angle ϕ of a rotated image with respect to its reference image with an angle θ, is the same projection, but scaled and translated, on the reference image at the angle $\phi + \theta$. Therefore, if all the transformation parameters are known, and assuming that the rotation angle is not zero that is, $\theta \neq 0$, then the $N + 1$ projections needed could be separated between the different channels, in order to sample and perfectly reconstruct the reference image in a symmetric manner.

The next question would be, how can we estimate the transformation parameters? We know that with the use of polynomial reproducing kernels, we can obtain the geometric moments of a signal, and geometric moments up to order 2 from two signals are enough to estimate translation, rotation and scaling parameters between the two signals. We also know that, as E-splines are a generalized version of B-splines [3], we can reproduce a combination of polynomials and exponentials from E-splines. From the polynomials moments up to order 2, we can estimate all the transformation parameters.

4. RESULTS

As an example, in [5] we showed that to achieve perfect reconstruction for a 4-sided bilevel polygon, a 2-D E-spline order of 12 is required to produce 5 projections at the angles $0, 45, 90, tan^{-1}(2)$ and $tan^{-1}(\frac{1}{2})$. With 2-D E-spline order of 7 however we can produce 3 projections at the angles $0, 45, 90$ on the reference signal, and a 2-D E-spline order of 7 on the second signal would give 3 projections for the reference signal at the angles $\theta, 45 + \theta, 90 + \theta$ where θ is the rotation parameter. Assuming θ is not zero, we would have enough projections to perfectly reconstruct the reference signal. Therefore an spline order of 7+2 = 9 (2 is needed for estimating the transformation parameters) on each signal would give us enough projections to perfectly reconstruct the reference signal. An example for a 4-sided bilevel polygon with two acquisition devices is shown in Figure 3 where the reference signal, its translated, rotated and scaled version, their samples, the E-spline sampling kernel, and the reconstructed reference signal are all shown.

5. CONCLUSION

In this paper we showed that with the use of Radon projections and generalized E-splines, symmetric multichannel sampling of translated, rotated and scaled bilevel polygons can be achieved. For estimating the geometrical transformations, we showed that as E-splines are a generalized version of B-splines, we can reproduce combination of polynomials and exponentials from E-splines. Therefore from the polynomial moments up to order 2, we can estimate all the unknown transformation parameters. For symmetric multichannel sampling of geometrically transformed bilevel polygons, we illustrated that the N+1 Radon projections needed for perfect reconstruction of an N-sided bilevel polygon, can be separated between the different channels, assuming that the rotation parameter is not zero. Our sampling and reconstruction algorithm is based on noise-free communication between the transmitter and receiver which is rather not very practical. The future research of this work is to test the stability and performance of our method in the presence of noise.

6. REFERENCES

[1] M. Vetterli, P. Marziliano and T. Blu, "Sampling Signals with Finite Rate of Innovation", IEEE Transactions on Signal Processing, vol. 50, pp. 1417-1428, June 2002.

[2] P.L. Dragotti, M. Vetterli and T. Blu, "Sampling Moments and Reconstructing Signals of Finite Rate of Innovation: Shannon meets Strang-Fix", IEEE Transactions on Signal Processing, vol. 55, pp. 1741-1757, May 2007.

[3] M. Unser and T. Blu, "Cardinal Exponential Splines: Part I - Theory and Filtering Algorithms", IEEE Transactions on Signal Processing, vol. 53, pp. 1425, 2005.

[4] L. Baboulaz, "Feature Extraction for Image Super-resolution using Finite Rate of Innovation Principles", PhD thesis, Department of Electrical and Electronic Engineering, Imperial College London, 2008. URL: http://www.commsp.ee.ic.ac.uk/~lbaboula/

[5] H. Akhondi Asl and P.L. Dragotti, "Single and Multichannel Sampling of Bilevel Polygons Using Exponential Splines", To Appear on IEEE International Conference on Acoustics, Speech, and Sig-

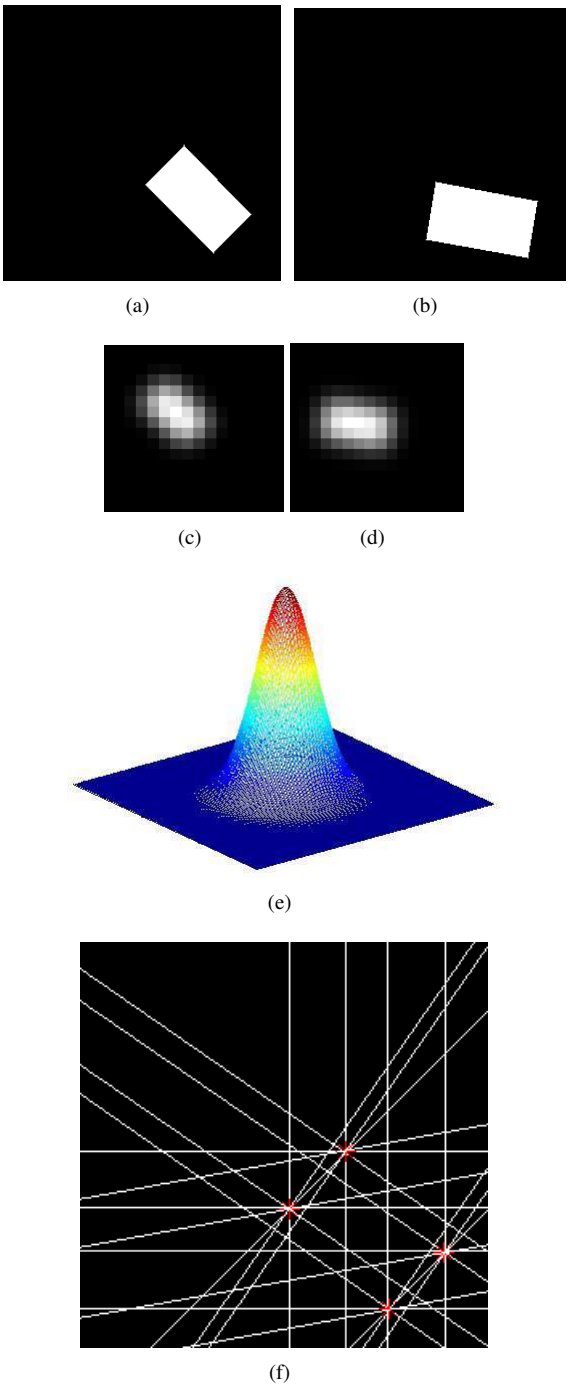


Fig. 3. **Symmetric multichannel sampling of translated, rotated and scaled bilevel polygons using E-spline sampling kernels. (a) The reference signal in a frame data size of 256×256. (b) The translated $(\triangle \mathbf{x} = -\mathbf{100}, \triangle \mathbf{y} = \mathbf{150})$, rotated $(\theta = \mathbf{35})$ and scaled $(\mathbf{a} = \mathbf{1.1})$ version of the reference signal. (c) & (d) The 16×16 samples of both signals. (e) 2-D generalized E-spline of order 9 (f) The reconstructed vertices of the reference signal with 6 back-projections, the crosses are the actual vertices of the polygon. [Not to scale]**

nal Processing, Taipei, Taiwan, April 2009. URL: http://cspserver2.ee.ic.ac.uk/~Hojakndi/

[6] C. S. Seelamantula and M. Unser, “A Generalized Sampling Method for Finite-Rate-of-Innovation-Signal Reconstruction”, IEEE Signal Processing Letters, vol.15, pp. 813-816, August 2008.

[7] J. Kusuma and V. K. Goyal, "Multichannel Sampling of Parametric Signals with a Successive Approximation Property," IEEE International Conference on Image Processing, pp. 1265-1268, October 2006.

[8] L. Baboulaz and P. L. Dragotti, "Distributed Acquisition and Image Super-Resolution Based on Continuous Moments from Samples," IEEE International Conference on Image Processing, pp. 3309-3312, October 2006.

[9] I. Maravic and M. Vetterli, “Exact sampling results for some classes of parametric non-bandlimited 2-D signals", IEEE Transactions on Signal Processing, vol.52, no.1, pp. 175-189, January 2004ation Principles”, PhD thesis, Imperial College London, 2008.

[10] G. T. Herman, “Image Reconstruction from Projections: The Fundamentals of Computerized Tomography”, Academic Press, New York, 1980.

[11] F. Vanpoucke, M. Moonen and Y. Berthoumieu, “An Efficient Subspace Algorithm for 2-D Harmonic Retrieval", IEEE International Conference on Acoustics, Speech, and Signal Processing, vol.4, pp. 461-464, April 1994.

[12] P. Shukla and P.L. Dragotti, “Sampling Schemes for Multidimensional Signals with Finite Rate of Innovation”, IEEE Transactions on Signal Processing, vol. 55, pp. 3670-3686, July 2007.

[13] I. Maravic and M. Vetterli, “Exact sampling results for some classes of parametric non-bandlimited 2-D signals", IEEE Transactions on Signal Processing, vol.52, no.1, pp. 175-189, January 2004.

[14] H. Akhondi Asl, “Single and Multichannel Sampling of Signals With Finite Rate of Innovation Using E-Splines”, MPhil to PhD Transfer Report, Department of Electrical and Electronic Engineering, Imperial College London, 2008. URL: http://cspserver2.ee.ic.ac.uk/~Hojakndi/

Special session on

Sampling and Quantization

Chair: Özgür Yilmaz

Quantization for Compressed Sensing Reconstruction

John Z. Sun and Vivek K Goyal

Massachusetts Institute of Technology, Cambridge, MA 02139 USA
johnsun@mit.edu, vgoyal@mit.edu

Abstract:

Quantization is an important but often ignored consideration in discussions about compressed sensing. This paper studies the design of quantizers for random measurements of sparse signals that are optimal with respect to mean-squared error of the lasso reconstruction. We utilize recent results in high-resolution functional scalar quantization and homotopy continuation to approximate the optimal quantizer. Experimental results compare this quantizer to other practical designs and show a noticeable improvement in the operational distortion-rate performance.

1. Introduction

In practical systems where information is stored or transmitted, data must be discretized using a quantization scheme. The design of the optimal quantizer for a given stochastic source has been well studied and is surveyed in [6]. Here, optimal means the quantizer minimizes the error as measured by some distortion metric. In this paper, we explore optimal quantization for an emerging nonadaptive compression paradigm called compressed sensing (CS) [1, 4]. Several authors have studied the asymptotic reconstruction performance of quantized random measurements assuming a mean-squared error (MSE) distortion metric [3, 5]. Other previous work presented modifications to existing reconstruction algorithms to mitigate distortion resulting from standard quantizers [3, 7] or modified quantization that can be viewed as the binning of quantizer output indexes [10].

Our contribution is to reduce distortion due to quantization through design of the quantizer itself. The key observation is simply that the random measurements are used as arguments in a *nonlinear* reconstruction function. Thus, minimizing the MSE of the measurements is not equivalent to minimizing the MSE of the reconstruction. We use the theory for high-resolution distributed functional scalar quantization (DFSQ) recently developed in [9] to design optimal quantizers for random measurements. To obtain concrete results, we choose a particular reconstruction function (lasso [11]) and distributions for the source data and sensing matrix. However, the general principle of obtaining improvements through the use of DFSQ theory holds more generally, and we address the conditions that must be satisfied for sensing and reconstruction. Also, rather than develop results for fixed and variable rate in parallel, we present only fixed rate. To concentrate on the central ideas, we choose signal and sensing models that obviate discussion of quantizer overload.

2. Background

In our notation, a random vector is always lowercase and in bold. A subscript then indicates an element of the vector. Also, an unbolded vector y corresponds to a realization of the random vector $\mathbf{y}$.

2.1 Distributed functional scalar quantization

In standard fixed-rate scalar quantization [6], one is asked to design a quantizer Q that operates separably over its components and minimizes MSE between a probabilistic source vector $\mathbf{y} \in \mathbb{R}^M$ and its quantized representation $\hat{\mathbf{y}} = Q(\mathbf{y})$. The resulting optimization is

$$\min_Q E\left[\|\mathbf{y} - Q(\mathbf{y})\|^2\right],$$

subject to the constraint that the maximum number of codewords or quantization levels for each $\mathbf{y_i}$ is less than 2^{R_i}. We can use high-resolution theory to find the quantizer point density of the optimal quantizer.

In DFSQ [9], the goal is to create a quantizer that minimizes distortion for some scalar function $g(\mathbf{y})$ of the source vector $\mathbf{y}$ rather than the vector itself. Hence, the optimization is now

$$\min_Q E\left[|g(\mathbf{y}) - g(Q(\mathbf{y}))|^2\right]$$

such that the maximum number of codewords or quantization levels representing each $\mathbf{y_i}$ is less than 2^{R_i}. To apply the following model, we need $g(\cdot)$ and $f_{\mathbf{y}}(\cdot)$ to satisfy certain conditions:

C1. $g(\mathbf{y})$ is smooth and monotonic for each $\mathbf{y_i}$.

C2. The partial derivative $g_i(y) = \partial g(y)/\partial y_i$ is defined and bounded for each i.

C3. The joint pdf of the source variables $f_{\mathbf{y}}(y)$ is smooth and supported in a compact subset of $\mathbb{R}^M$.

For valid $g(\cdot)$ and $f_{\mathbf{y}}(\cdot)$ pairs, we define a set of functions

$$\gamma_i(t) = \left(E\left[|g_i(\mathbf{y})|^2 \mid \mathbf{y_i} = t\right]\right)^{1/2}. \quad (1)$$

We call $\gamma_i(t)$ the *sensitivity* of $g(\mathbf{y})$ with respect to the source variable $\mathbf{y_i}$. The optimal point density is then

$$\lambda_i(t) = C\left(\gamma_i^2(t) f_{\mathbf{y_i}}(t)\right)^{1/3}, \quad (2)$$

for some normalization constant C, which leads to a total operational distortion-rate

$$D(\{R_i\}) = \sum_i 2^{-2R_i} E\left[\frac{\gamma_i^2(\mathbf{y_i})}{12\lambda_i^2(\mathbf{y_i})}\right]. \qquad (3)$$

The sensitivity $\gamma_i(t)$ serves to reshape the quantizer, giving better resolution to regions of $\mathbf{y_i}$ that have more impact on $g(\mathbf{y})$, thereby reducing MSE.
Similar results for variable-rate quantizers are also presented in [9]. However, we will only consider the fixed-rate case in this paper. The theory of DFSQ can be extended to a vector of functions, where $\mathbf{x_j} = g^{(j)}(\mathbf{y})$ for $1 \le j \le N$. Since the cost function is additive in its components, we can show that the overall sensitivity for each component $\mathbf{y_i}$ is

$$\gamma_i(t) = \frac{1}{N}\sum_{j=1}^{N} \gamma_i^{(j)}(t), \qquad (4)$$

where $\gamma_i^{(j)}(t)$ is the sensitivity of the function $g^{(j)}(\mathbf{y})$ with respect to $\mathbf{y_i}$.

2.2 Compressed Sensing

CS refers to estimation of a signal at a resolution higher than the number of data samples, taking advantage of sparsity or compressibility of the signal and randomization in the measurement process [1, 4]. We will consider the following formulation. The input signal $x \in \mathbb{R}^N$ is K-sparse in some orthonormal basis Ψ, meaning the transformed signal $u = \Psi^{-1}x \in \mathbb{R}^N$ contains only K nonzero elements. Consider a length-M measurement vector $y = \Phi x$, where $\Phi \in \mathbb{R}^{M\times N}$ with $K < M < N$ is a realization of $\mathbf{\Phi}$. The major innovation in CS (for the case of sparse u considered here) is that recovery of x from y via some computationally-tractable reconstruction method can be guaranteed asymptotically almost surely.
Many reconstruction methods have been proposed including a linear program called basis pursuit [2] and greedy algorithms like orthogonal matching pursuit (OMP) [12]. In this paper, we focus on a convex optimization called lasso [11], which takes the form

$$\hat{x} = \arg\min_x \left(\|y - \Phi x\|_2^2 + \mu\|\Psi^{-1}x\|_1\right). \qquad (5)$$

As one sample result, lasso leads to perfect sparsity pattern recovery with high probability if $M \sim 2K\log(N-K) + K$ under certain conditions on $\mathbf{\Phi}$, μ, and the scaling of the smallest entry of u [13]. Unlike in [5], our concern in this paper is not how the scaling of M affects performance, but rather how the accuracy of the lasso computation (5) is affected by quantization of y.
A method for understanding the set of solutions to (5) is the homotopy continuation (HC) method [8]. HC considers the regularization parameter μ at an extreme point (e.g., very large μ so the reconstruction is all zero) and slowly varies μ so that all sparsities and the resulting reconstructions are obtained. It is shown that there are N values of μ where the lasso solution changes sparsity, or equivalently $N+1$ intervals over which the sparsity does not change. For μ in the interior of one of these intervals, the reconstruction is determined uniquely by the solution of a linear system of equations involving a submatrix of Φ. In particular, for a specific choice μ^* and observed random measurements y,

$$2\Phi_{J_{\mu^*}}^T \Phi_{J_{\mu^*}} \hat{x} + \mu^* v = 2\Phi_{J_{\mu^*}}^T y, \qquad (6)$$

where $v = \mathrm{sgn}(\hat{x})$ and $\Phi_{J_{\mu^*}}$ is the submatrix of Φ with columns corresponding to the nonzero elements $J_{\mu^*} \subset \{1, 2, \ldots, N\}$ of $\hat{x}$.

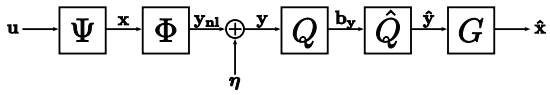


Figure 1: A compressed sensing model with quantization of measurement vector $\mathbf{y}$. The vector $\mathbf{y_{nl}}$ denotes the noiseless random measurements.

3. Problem Model

Figure 1 presents a CS model with quantization. Assume without loss of generality that $\Psi = I_N$ and hence the (random) signal $\mathbf{x} = \mathbf{u}$ is K-sparse. Also assume a random matrix $\mathbf{\Phi}$ is used to take measurements, and additive Gaussian noise perturbs the resulting signal, meaning the continuous-valued measurement vector is $\mathbf{y} = \mathbf{\Phi x} + \boldsymbol{\eta}$. The sampler wants to transmit the measurements with total rate R and encodes $\mathbf{y}$ into a transmittable bitstream $\mathbf{b_y}$ using encoder Q. Next, a decoder $\hat{Q}$ produces a quantized signal $\hat{\mathbf{y}}$ from $\mathbf{b_y}$. Finally, a reconstruction algorithm G outputs an estimate $\hat{\mathbf{x}}$. The function G is a black box that may represent lasso, OMP or another CS reconstruction algorithm.
We now present a probabilistic model for the input source and sensing matrix. It is chosen to guarantee finite support on both the input and measurement vectors, and prevent overload errors for quantizers with small R. However, we emphasize that the following theory is general, and other choices for $\mathbf{x}$ and $\mathbf{\Phi}$ are possible for large enough R.
Assume the K-sparse vector $\mathbf{x}$ has random sparsity $\mathbf{J}$ chosen uniformly from all possibilities, and each nonzero component $\mathbf{x_i}$ is distributed iid $\mathcal{U}(-1,1)$. Also assume the additive noise vector $\boldsymbol{\eta}$ is distributed iid Gaussian with zero mean and variance σ^2. Finally, let $\mathbf{\Phi}$ correspond to random projections such that each column $\boldsymbol{\phi_j} \in \mathbb{R}^M$ has unit energy ($\|\boldsymbol{\phi_j}\|^2 = 1$). The columns of $\mathbf{\Phi}$ thus form a set of N random vectors chosen uniformly on the unit $(M-1)$-hypersphere. Since $\mathbf{y} = \Phi\mathbf{x}$,

$$\mathbf{y_i} = \sum_{j=1}^{N} \mathbf{\Phi_{ij}x_j} = \sum_{j\in\mathbf{J}} \underbrace{\mathbf{\Phi_{ij}x_j}}_{\mathbf{z_{ij}}}.$$

The distribution of each $\mathbf{z_{ij}}$ is found using derived distributions. The resulting pdfs can be shown to be iid $f_{\mathbf{z}}(z)$, where $\mathbf{z}$ is a scalar random variable that is identical in distribution to each $\mathbf{z_{ij}}$. The distribution of $\mathbf{y_i}$ is then the $K-1$ convolution cascade of $f_{\mathbf{z}}(z)$ with itself. Thus, $f_{\mathbf{y}}(y)$ is smooth and supported for $\{|\mathbf{y_i}| \le K\}$, satisfying

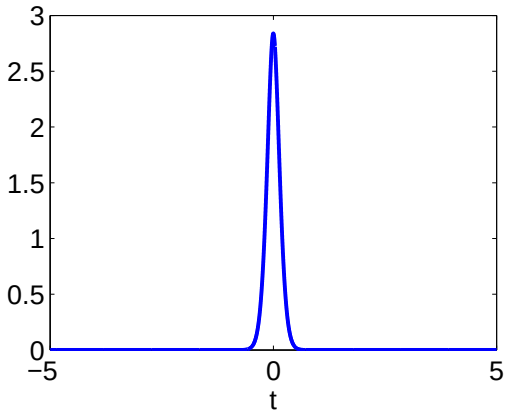


Figure 2: Distribution $f_{\mathbf{y_i}}(t)$ for $(K, M, N) = (5, 71, 100)$. The support of $\mathbf{y_i}$ is the range $[-K, K]$, where K is the sparsity of the input signal. However, the probability is only non-negligible for small $\mathbf{y_i}$.

condition C3 for DFSQ. Figure 2 illustrates the distribution of $\mathbf{y_i}$ for a particular case.

The reconstruction algorithm G is a function of the measurement vector $\mathbf{y}$ and sampling matrix $\mathbf{\Phi}$. We will show that if $G(\mathbf{y}, \mathbf{\Phi})$ is lasso with a proper relaxation variable μ, then conditions C1 and C2 are met. Using HC, we see $G(\mathbf{y}, \mathbf{\Phi})$ is a piecewise smooth function that is also piecewise monotonic with every $\mathbf{y_i}$ for a *fixed* μ. Moreover, for every μ the reconstruction is an affine function of the measurements through (6), so the partial derivative with respect to any element y_i is piecewise defined and smooth (constant in this case). Conditions C1 and C2 are therefore satisfied.

4. Optimal Quantizer Design

We now pose the optimal fixed-rate quantizer design as a DFSQ problem. For a given noise variance σ^2, choose an appropriate μ^* to form the best reconstruction $\hat{\mathbf{x}}$ from the unquantized random measurements $\mathbf{y}$. We produce M quantizers to transmit the elements of $\mathbf{y}$ such that the decoded message $\hat{\mathbf{y}}$ will minimize the distortion between $\tilde{\mathbf{x}} = G(\mathbf{y}, \mathbf{\Phi})$ and $\hat{\mathbf{x}} = G(\hat{\mathbf{y}}, \mathbf{\Phi})$ for a total rate R. Note G can be visualized as a set of N scalar functions $\hat{\mathbf{x}}_\mathbf{j} = G^{(j)}(\hat{\mathbf{y}}, \mathbf{\Phi})$ that are identical in distribution due to symmetry in the randomness of $\mathbf{\Phi}$. Since the sparse input signal is assumed to have uniformly distributed sparsity and $\mathbf{\Phi}$ distributes energy uniformly to all measurements $\mathbf{y_i}$ in expectation, we argue by symmetry that each measurement is allotted the same number of bits and that every measurement's quantizer is the same. Moroever, since the functions representing the reconstruction are identical, we argue using (4) that the overall sensitivity $\gamma_{\text{cs}}(\cdot)$ is the same as the sensitivity of any $G^{(j)}(\hat{\mathbf{y}}, \mathbf{\Phi})$. Computing (2) yields the point density $\lambda_{\text{cs}}(\cdot)$.

This is when the homotopy continuation method becomes extremely useful. For a given realization of $\mathbf{\Phi}$ and $\boldsymbol{\eta}$, we can use HC to determine how many elements in the reconstruction are nonzero for μ^*, denoted J_{μ^*}. Equation (6) is then used to find $\partial G^{(j)}(y, \Phi)/\partial y_i$, which is needed to compute $\gamma_{\text{cs}}(\cdot)$. To simplify our notation, let $A = \Phi_{J_{\mu^*}}$. The resulting differentials can be expressed as

$$\frac{\partial G^{(j)}(y, \Phi)}{\partial y_i} = \left[\left(A^T A\right)^{-1} A^T \right]_{ji}. \tag{7}$$

We now present the sensitivity through the following theorem:

Theorem 1 *Let the noise variance be σ^2 and choose an appropriate μ^*. Define $\mathbf{y}_{\setminus \mathbf{i}}$ to be all the elements of a vector $\mathbf{y}$ except $\mathbf{y_i}$. The sensitivity of each element $\mathbf{y_i}$, which is denoted $\gamma_i^{(j)}(t)$, can be written as*

$$\left(E_{\mathbf{\Phi}, \mathbf{y}_{\setminus \mathbf{i}}} \left[\frac{f_{\mathbf{y_i}|\Phi}(t|\Phi)}{f_{\mathbf{y_i}}(t)} \left[\left(A^T A\right)^{-1} A^T \right]_{ji} \mid \mathbf{y_i} = t \right] \right)^{\frac{1}{2}},$$

where A is the submatrix of Φ as described in HC for μ^* and some observation $\mathbf{y}$. Moreover, for any Φ and its corresponding J, $f_{\mathbf{y_i}|\mathbf{\Phi}}(t|\Phi)$ is the convolution cascade of $\{\mathbf{z_j} \sim \mathcal{U}(-\Phi_{ij}, \Phi_{ij})\}$ for $j \in J$. By symmetry arguments, $\gamma_{\text{cs}}(t) = \gamma_i^{(j)}(t)$ for any i and j.

This expectation is difficult to calculate but can be approached through L Monte Carlo trials on $\mathbf{\Phi}$, $\boldsymbol{\eta}$, and $\mathbf{x}$. For each trial, we can compute the partial derivative using (7). We denote the Monte Carlo approximation to that function to be $\gamma_{\text{cs}}^{(L)}(\cdot)$. Its form is

$$\gamma_{\text{cs}}^{(L)}(t) = \frac{1}{L} \sum_{\ell=1}^{L} \left(\frac{f_{\mathbf{y_i}|\mathbf{\Phi}}(t|\Phi_\ell)}{f_{\mathbf{y_i}}(t)} \left[\left(A_\ell^T A_\ell\right)^{-1} A_\ell^T \right]_{ji}^2 \right)^{\frac{1}{2}}, \tag{8}$$

with i and j arbitrarily chosen. By the weak law of large numbers, the empirical mean of L realizations of the random parameters should approach the true expectation for L large.

We now substitute (8) into (2) to find the Monte Carlo approximation to the optimal quantizer for compressed sensing. It becomes

$$\lambda_{\text{cs}}^{(L)}(t) = C \left(\gamma_{\text{cs}}^{(L)}(t) f_{\mathbf{y_i}}(t) \right)^{1/3}, \tag{9}$$

for some normalization constant C. Again by the weak law of large numbers, $\lambda_{\text{cs}}^{(L)}(t) \xrightarrow{p} \lambda_{\text{cs}}(t)$ for L large.

5. Experimental Results

We compare the CS-optimized quantizer, called the "sensitive" quantizer, to a uniform quantizer and "ordinary" quantizer $\lambda_{\text{ord}}(t)$ which is optimized for the distribution of $\mathbf{y}$. This means the ordinary quantizer would be best if we want to minimize distortion between $\mathbf{y}$ and $\hat{\mathbf{y}}$, and hence has a flat sensitivity curve over the support of $\mathbf{y}$. The sensitive quantizer $\lambda_{\text{cs}}(t)$ is found using (9) and the uniform quantizer $\lambda_{\text{uni}}(t) = c$, where c is a normalization constant.

Using 1000 Monte Carlo trials, we estimate $\gamma_{\text{cs}}(t)$. The resulting point density functions for the three quantizers are illustrated in Figure 3.

Experimental results are performed on a Matlab testbench. Practical quantizers are designed by extracting codewords

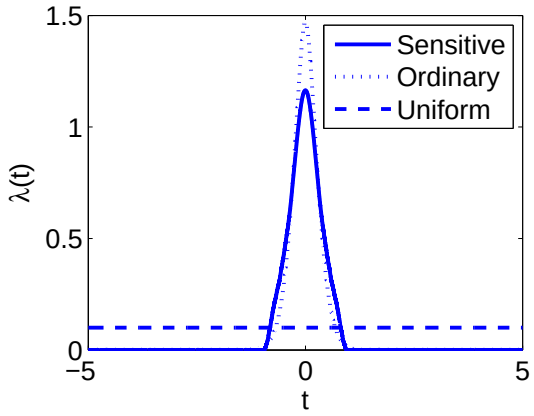


Figure 3: Estimated point density functions $\lambda_{\mathrm{cs}}(t)$, $\lambda_{\mathrm{ord}}(t)$, and $\lambda_{\mathrm{uni}}(t)$ for $(K, M, N) = (5, 71, 100)$.

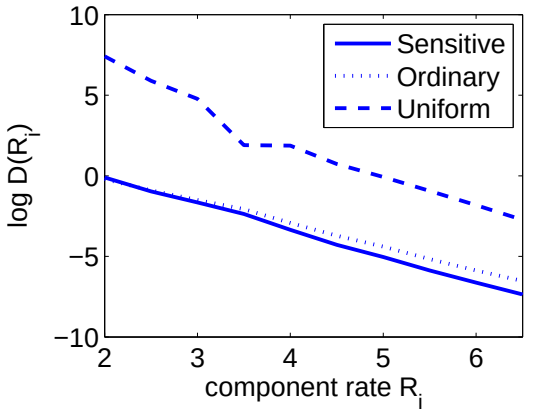


Figure 4: Results for distortion-rate for the three quantizers with $\mu = 0.01$ and $\sigma^2 = 0.3$. We see that the sensitive quantizer has the least distortion.

from the cdf of the normalized point densities. In the approximation, the ith codeword is the point t such that

$$\int_{-\infty}^{t} \lambda_{\mathrm{cs}}(t')dt' = \frac{i - 1/2}{2^{R_i}},$$

where R_i is the rate for each measurement. The partition points are then chosen to be the midpoints between codewords.

We compare the sensitive quantizer to uniform and ordinary quantizers using the parameters $\mu = 0.1$ and $\sigma^2 = 0.3$. Results are shown in Figure 4.

We find the sensitive quantizer performs best in experimental trials for this combination of μ and σ^2 at sufficiently high rates. This makes sense because $\lambda_{\mathrm{cs}}(t)$ is a high-resolution approximation and should not necessarily perform well at very low rates.

6. Conclusion

We present a high-resolution approximation to an optimal quantizer for the storage or transmission of random measurements in a compressed sensing system with lasso reconstruction. Using DFSQ and HC, we find a sensitivity function $\gamma_{\mathrm{cs}}(\cdot)$ that determines the optimal point density function $\lambda_{\mathrm{cs}}(\cdot)$ of such a quantizer. Experimental results show that the operational distortion-rate is best when using this so called "sensitive" quantizer.

We conclude that proper quantization in compressed sensing is not simply a function of the distribution of the random measurements themselves (using either a high-resolution approximation or practical algorithms like Lloyd-Max). Rather, quantization adds a non-constant effect, called functional sensitivity [9], on the distortion between the the lasso reconstructions of the random measurements and its quantized version.

A significant amount of work can still be done in this area. Parallel developments could be made for variable-rate quantizers. Also, this theory can be extended to other probabilistic signal and sensing models, and CS reconstruction methods.

References:

[1] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

[2] S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comp.*, 20(1):33–61, 1999.

[3] W. Dai, H. Vinh Pham, and O. Milenkovic. Quantized compressive sensing. arXiv:0901.0749v2 [cs.IT]., 2009.

[4] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[5] V. K. Goyal, A. K. Fletcher, and S. Rangan. Compressive sampling and lossy compression. *IEEE Sig. Process. Mag.*, 25(2):48–56, 2008.

[6] R. M. Gray and D. L. Neuhoff. Quantization. *IEEE Trans. Inform. Theory*, 44(6):2325–2383, 1998.

[7] L. Jacques, D. K. Hammond, and M. J. Fadili. Dequantized compressed sensing with non-Gaussian constraints. arXiv:0902.2367v2 [math.OC]., 2009.

[8] D. M. Malioutov, M. Cetin, and A. S. Willsky. Homotopy continuation for sparse signal representation. In *Proc. IEEE ICASSP*, pp. 733–736, 2006.

[9] V. Misra, V. K. Goyal, and L. R. Varshney. Distributed functional scalar quantization: High-resolution analysis and extensions. arXiv:0811.3617v1 [cs.IT]., 2008.

[10] R. J. Pai. Nonadaptive lossy encoding of sparse signals. Master's thesis, Massachusetts Inst. of Tech., Cambridge, MA, 2006.

[11] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Stat. Soc., Ser. B*, 58(1):267–288, 1996.

[12] J. A. Tropp. Greed is good: Algorithmic results for sparse reconstruction. *IEEE Trans. Inform. Theory*, 50(10):2231–2242, 2004.

[13] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity. *Department of Statistics, UC Berkley, Tech. Rep 709*, 2006.

Finite Range Scalar Quantization for Compressive Sensing

Jason N. Laska [(1)], Petros Boufounos [(2)], and Richard G. Baraniuk[(1)]

(1) Rice University, 6100 Main St., Houston, TX 77005
(2) Mitsubishi Electric Research Laboratories, 201 Broadway Cambridge, MA 02139
laska@rice.edu, petrosb@merl.com, richb@rice.edu

Abstract:

Analog-to-digital conversion comprises of two fundamental discretization steps: sampling and quantization. Recent results in compressive sensing (CS) have overhauled the conventional wisdom related to the sampling step, by demonstrating that sparse or compressible signals can be sampled at rates much closer to their sparsity rate, rather than their bandwidth. This work further overhauls the conventional wisdom related to the quantization step by demonstrating that quantizer overflow can be treated differently in CS and by exploiting the tradeoff between quantization error and overflow.
We demonstrate that contrary to classical approaches that avoid quantizer overflow, a better finite-range scalar quantization strategy for CS is to amplify the signal such that the finite range quantizer overflows at a pre-determined rate, and subsequently reject the overflowed measurements from the reconstruction. Our results further suggest a simple and effective automatic gain control strategy which uses feedback from the saturation rate to control the signal gain.

1. Introduction

Analog-to-digital converters (ADCs) are an essential part of most modern sensing and communications systems. They are the interface between the analog physical world and the digital processing world that extracts the information we are interested in. Ever-increasing demands for information has pushed the requirements on ADCs to their current physical limits. Fortunately, recent theoretical developments in the area of compressive sensing (CS) enable us to significantly extend the capabilities of current ADCs to keep pace with demand.
CS is a framework that allows signals that have sparse representation, i.e., few non-zero elements, or few non-zero coefficients in some basis, to be sampled at a rate close to the sparsity rate, rather than the Nyquist rate. CS employs linear measurement systems and a non-linear reconstruction algorithms to acquire and recover sparse signals.
Most of the CS literature to-date focuses on one particular aspect of ADCs, namely sampling. In this paper we reexamine the other significant aspect, quantization. Specifically, we show that the core tenets of CS enable us to reduce the error due to quantization by allowing the quantizer to saturate more often than usual and removing the saturated measurements from the reconstruction process.
The organization of this paper is as follows. Section 2. presents a brief background on analog-to-digital conversion, compressive sampling, and finite-range quantization. Section 3. presents a brief analysis of finite-range quantization for CS. We show that CS measurements and the quantization error are i.i.d. Gaussian, and analyze the proposed reconstruction strategy. Section 4., presents numerical results that validate our analysis. We conclude with a brief discussion in Sec. 5.

2. Background

2.1 Analog-to-digital conversion

Analog-to-digital conversion consists of two discretization steps: *sampling*, which converts an analog signal to a set of discrete measurements, and *quantization*, which converts each real-valued measurement to a discrete one chosen from a pre-determined set. Although both steps are necessary to represent a signal in the discrete digital world, classical results due to Shannon and Nyquist demonstrate that the sampling step is information preserving if a sufficient number of samples, i.e., measurements, are obtained. On the other hand quantization always degrades the signal. The system design to goal is to take enough measurements such that the signal does not alias, and to acquire enough bits to limit the quantization distortion.

2.2 Finite-range quantization

Scalar quantization is the process of converting the continuous value of the measurements to one of several discrete values through a non-invertible function $R(\cdot)$. In this paper we focus on uniform quantizers with quantization interval Δ. Thus, the quantization points are $q_k = q_0 + k\Delta$, and every scalar a is quantized to the nearest quantization point $R(a) = \operatorname{argmin}_{q_k} |a - q_k|$. For an infinite-range quantizer this implies that the quantization error is bounded by $|a - R(q)| \leq \Delta/2$.
In practice quantizers have finite range, dictated by hardware constraints such as the voltage limits of the devices and the finite bit-rate of the quantized representation. Without loss of generality we assume a midrise B-bit quantizer that represents a symmetric range of values $|a| < T$, where $T > 0$ is the quantization threshold. The corresponding quantization points are at $q_k =$

$\Delta/2 + k\Delta, k = -2^{B-1}, \ldots, 2^{B-1} - 1$. This assumption implies a quantization interval $\Delta = 2^{-B+1}T$. Any measurement with magnitude greater than T saturates the quantizer and "clips" to magnitude T, i.e., it quantizes to the quantization point $T - \Delta/2$.

Most classical quantization error analysis assumes that the measurements are scaled such that the quantizer never clips. This is a sensible quantization strategy for classical approaches using linear reconstruction. In that context, saturation events cause significant signal distortion and are undesirable. For that reason, extreme attention is often devoted to pre-ADC automatic gain control (AGC) systems to ensure that the quantizer saturates only rarely. Under this assumption the analysis of a finite or an infinite range quantizer is equivalent in terms of the quantization error. Thus, an infinite-range quantizer is often assumed for its mathematical simplicity.

2.3 Compressive sampling (CS)

The theory of *compressive sampling* (CS) overhauls the conventional wisdom on the sampling process. Specifically, [2] and the references therein show that the number of measurements that are sufficient to exactly reconstruct a sampled signal are significantly fewer than the Shannon-Nyquist rate as long as the signal is sparse, i.e., can be represented with very few non-zero components in some basis.

The key components of CS are *randomized measurements* and *non-linear reconstruction*. Specifically, a Nyquist-rate sampled discrete-time signal $\mathbf{x}$ can be sampled at a lower rate by using a random matrix Φ, of dimension $M \times N$:

$$\mathbf{y} = \Phi\mathbf{x}, \tag{1}$$

and reconstructed exactly, if the signal is K-*sparse*, i.e., only has K non-zero components in some basis and the matrix Φ satisfies the *Restricted Isometry Property* (RIP) [2]:

$$\sqrt{1-\delta_{2K}}\|\mathbf{x}\|_2 \leq \|\Phi\mathbf{x}\|_2 \leq \sqrt{1+\delta_{2K}}\|\mathbf{x}\|_2 \tag{2}$$

for all $2K$-sparse signals $\mathbf{x}$, where δ_{2K} is the RIP constant of Φ. RIP guarantees that the norm of the measurements does not deviate significantly from the norm of the K-sparse signal $\mathbf{x}$.

To reconstruct $\widehat{\mathbf{x}}$ from $\mathbf{y}+\mathbf{n}$, where $\mathbf{n}$ is noise with $\|\mathbf{n}\|_2 = \eta$, we perform the optimization

$$\widehat{\alpha} = \min_{\alpha} \|\alpha\|_1 \text{ s.t. } \|\Phi\Psi\alpha - \mathbf{y}\|_2 < \eta, \quad \widehat{\mathbf{x}} = \Psi\widehat{\alpha} \tag{3}$$

where Ψ is a basis and $\|\alpha\|_1 = \sum_i |\alpha_i|$ is the ℓ_1 norm of the coefficient vector. Reconstructing using (3) guarantees that the norm of the reconstruction error is bounded by $c\eta$, where c is a system-dependent constant [2].

In this paper we use the two key components of CS, namely randomized measurements and non-linear reconstruction, to overhaul the conventional wisdom on scalar quantization. In the next sections we demonstrate that the CS measurement process makes the quantization error a white noise process. We use that result demonstrate that in the context of non-linear reconstruction it is advantageous to scale the signal such that the quantizer saturates at a positive rate and reject the saturated measurements from the reconstruction.

3. Finite-range quantization for CS

The non-linear reconstruction methods used in CS and the *democratic* nature of the measurements, suggests that with only a small performance penalty, we can choose to ignore measurements. Specifically, in this work we choose to deliberately saturate the quantizer and ignore the measurements that saturated. In the analysis that follows we demonstrate the advantages of this approach compared to scaling the measurements such that they do not saturate or incorporating the saturated measurements in the reconstruction.

The analysis is based on three distinct results:

1. CS measurements approximately follow an i.i.d. Gaussian distribution, making the quantization error a well characterized white noise process.
2. Clipping without quantization followed by dropping the saturated measurements preserves the signal norm and the RIP.
3. Once quantization is introduced, the signal-to-quantization noise ratio can be minimized by selecting a positive saturation rate and rejecting the saturated measurements.

The subsequent sections state and sketch the proofs for these results and their consequences. Due to space limitations, we defer complete proofs and extended analysis to future publications.

3.1 Distribution of CS measurements

We assume the measurement matrix Φ in (1) is randomly generated using a zero-mean sub-Gaussian distribution with variance $1/M$. Under this assumption, all the measurements $y_i = \sum_j (\Phi)_{i,j} x_j$ are i.i.d. zero-mean random variables with variance $\|\mathbf{x}\|_2^2/M$. Using the Lyapunov variant of the Central Limit Theorem, it is also straightforward to show that as the dimension N of the signal $\mathbf{x}$ increases the y_i become normally distributed. The statement becomes non-asymptotic if the elements of Φ are themselves distributed as a Gaussian. Our initial experiments show that commonly used CS matrix families reach asymptotic behavior even for small N.

The implications of this statement are threefold:

1. The expected number of measurements exceeding in magnitude a threshold $T\|\mathbf{x}\|_2/\sqrt{M}$ is $2Q(T)$, where $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^{+\infty} e^{-t^2/2}dt$ is the tail integral of the standard Gaussian distribution.
2. The ratio of $T\|\mathbf{x}\|_2/\sqrt{M}$ determines the saturation rate. Thus, scaling the signal such that a specific saturation rate is achieved provides a very effective gain control strategy.
3. The quantization error is a white process, although it is correlated to the measurements.

We should note that in the sequel only the ratio $T\sqrt{M}/\|\mathbf{x}\|_2$ is relevant. This ratio is the threshold we select by varying the parameter T. The $\sqrt{M}$ factor reflects that in practical systems the variance of the elements of the measurement matrix is not a function of the number of measurements. The normalization by $\|\mathbf{x}\|_2$ reflects that in practice automatic gain control or prior signal knowledge is used to determine the proper gain in the input.

3.2 Analysis of finite-range CS measurements

In this section we introduce clipping at threshold $T\|x\|_2/\sqrt{M}$, without quantization. We reject the clipped measurements and demonstrate that if the remaining measurements, denoted using $\widetilde{\mathbf{y}}$, are sufficient in number, the measurement process still satisfies the RIP and preserves the norm of K-sparse signals. We use the notation $\widetilde{(\cdot)}$ to denote the relevant quantities after the saturated measurements are dropped: $\widetilde{M}$ is the number of remaining measurements and $\widetilde{\Phi}$ the mutilated measurement matrix corresponding to the remaining measurements.

Assuming the result of Sec. 3.1, the expected number of saturated measurements is $2MQ(T)$. The remaining $\widetilde{M}$ measurements follow a truncated Gaussian distribution:

$$\widetilde{y_i} \propto \begin{cases} \mathcal{N}\left(y_i; 0, \frac{\|x\|_2^2}{M}\right), & |y_i| < \frac{T\|\mathbf{x}\|_2^2}{\sqrt{M}} \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

Thus, the expected norm of $\widetilde{\mathbf{y}}$ is equal to:

$$E\{\|\widetilde{\mathbf{y}}\|_2^2\} = M(1-2Q(T))\sigma_T^2, \quad (5)$$

where σ_T^2 is the variance of (4). Thus, the scaled system

$$G\widetilde{\mathbf{y}} = G\widetilde{\Phi}\mathbf{x} \quad (6)$$

$$G = \left(\frac{\|\mathbf{x}\|_2^2}{M(1-2Q(T))\sigma_T^2}\right)^{1/2} \quad (7)$$

$$= \left(\frac{\sqrt{2\pi}}{\sqrt{2\pi}(1-2Q(T)) - 2Te^{-T^2/2}}\right)^{1/2} \quad (8)$$

preserves the expected value of the norm of the signal. It is also straightforward to demonstrate that the density of the norm of the signal concentrates around its expected value with very high probability, in manner similar to [1, 3].

It is also possible to demonstrate that the resulting $\widetilde{\Phi}$, which is now signal-dependent, preserves the RIP for all K-sparse signals, as long as $\widetilde{M} = O(K\log(N/K))$, or equivalently $M = O(K\log(N/K)/(1-2Q(T))$. The proof is beyond the scope of this paper [5]. However, it is important since it guarantees recovery of the signal, and the robustness to noise we need in the next section.

3.3 Quantization noise

In this section we quantize the thresholded measurements using quantization interval $\Delta = 2^{-B+1}T\|\mathbf{x}\|_2/\sqrt{M}$:

$$R(\widetilde{\mathbf{y}}) = \widetilde{\mathbf{y}} + \widetilde{\epsilon_Q}, \quad (9)$$

where $\widetilde{\epsilon_Q}$ is the vector of the quantization error. From the results of Sec. 3.1 and the distribution of the measurements after thresholding it follows that ϵ_Q is a white random vector with elements distributed as a wrapped truncated Gaussian random variable and bounded by $\pm\Delta/2$. For small quantization intervals the distribution is well approximated by a uniform distribution in the same interval, with variance $\Delta^2/12$ [6]. Assuming a unit norm input $\mathbf{x}$ the expected squared norm of the quantization error is:

$$E\{\|\widetilde{\epsilon_Q}\|_2^2\} = M(1-2Q(T))\Delta^2/12 \quad (10)$$

$$= 2^{-2B}(1-2Q(T))T^2/3. \quad (11)$$

It can also be shown that for large M the measure of this norm concentrates around its mean. When properly scaled with the G in (8), the quantization error becomes:

$$E\{\|G\widetilde{\epsilon_Q}\|_2^2\} = \frac{\sqrt{2\pi}2^{-2B}}{3}\frac{T^2}{\sqrt{2\pi} - \frac{Te^{-T^2/2}}{(1-2Q(T))}}, \quad (12)$$

which suggests an optimal threshold T that minimizes the error.

If the RIP is guaranteed, the norm of reconstruction error can be bounded by $c\|G\widetilde{\epsilon_q}\|_2^2$ with very high probability [2]. For most practical applications, the minimizing T in (12) is not sufficient to guarantee RIP, and therefore we select the smallest T that does.

A similar analysis can be performed if we keep all the saturated measurements. In this case the RIP always holds and the measurement error is equal to:

$$E\{\|\epsilon_Q\|_2^2\} = \quad (13)$$

$$= M\left((1-2Q(T))\frac{\Delta^2}{12} + \frac{2Q(T)\|\mathbf{x}\|_2^2}{M}\sigma_{\text{trunc}}^2\right), \quad (14)$$

$$= \|\mathbf{x}\|_2^2\left((1-2Q(T))\frac{2^{-2B}}{3} + 2Q(T)\sigma_{\text{trunc}}^2\right), \quad (15)$$

where σ_{trunc}^2 is the variance of the tail distribution for a standard Gaussian random variable, as truncated by the saturation. Detailed analysis of this can be found in [4]. At T decreases, both σ_{trunc} and $Q(T)$ increases, which means the error due to the saturated measurements increases at the error due to the unsaturated measurements decreases. The optimal T in this case minimizes (15).

The two strategies can be compared to select the optimal given the operating conditions. Especially in low-bit conditions, reducing the quantization interval pays off in terms of the error. However, the tail effects cause a significant penalty if we keep the measurements, and the better strategy is to discard them. As we discuss in the next section in our extensive simulations under a large variety of practical conditions discarding the measurements performs better than using them.

4. Experimental validation

4.1 Experimental setup

Signal model: We study the performance of our approach using signals sparse in the frequency domain: in each trial K non-zero Fourier coefficients α_n are drawn from an i.i.d. Gaussian distribution, normalized to have unit norm, and randomly assigned to K frequency bins out of the N-dimensional space. The sampled signal $\mathbf{x}$ is the DFT of the generated Fourier coefficients. Beyond quantization we do not include additional noise sources. In addition to exactly sparse signals, we have performed extensive simulations with compressible signals and confirmed similar results. However, compressible signals are beyond the scope of this paper.

Measurement matrix: For each trial a measurement matrix is generated using a Rademacher distribution: each element is drawn independently to be $+1$ or -1 with equal probability. Our extended experimentation, not shown here in the interest of space, shows that our results are robust to large variety of measurement matrix classes.

Reconstruction metric: We report the reconstruction *signal-to-noise ratio* (SNR) in decibels (dB):

$$\text{SNR} \triangleq 10\log\left(\frac{\|\mathbf{x}\|_2^2}{\|\mathbf{x}-\widehat{\mathbf{x}}\|_2^2}\right), \quad (16)$$

where $\widehat{\mathbf{x}}$ denotes the reconstructed signal.

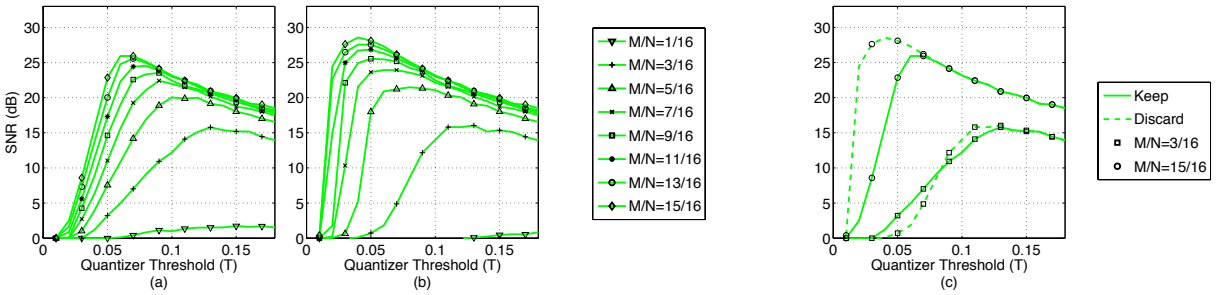


Figure 1: Reconstruction SNR (dB) vs. quantizer saturation threshold (T) using a 4-bit quantizer and downsampling rate $\frac{M}{N} = \frac{1}{16} \dots \frac{13}{16}$ when (a) the saturated measurements are used for reconstruction and (b) the saturated measurements are discarded before reconstruction. (c) Side-by-side comparison of (a) and (b) for $\frac{M}{N} = \frac{3}{16}$ and $\frac{15}{16}$: by lowering the threshold T and rejecting saturated measurements, we achieve the highest reconstruction SNR.

4.2 Experimental results

We performed extensive simulations with a variety of signal parameters. Due to space limitations, we present here the results for $N = 2048$, $K = 60$, and $B = 4$ which are typical of the system performance. In our experiments we vary M such that $\frac{M}{N} = \frac{1}{16} \dots \frac{15}{16}$ and the threshold T in the range $[0, 0.18]$. For each parameter combination we repeat 100 trials, each trial with a different signal $\mathbf{x}$ and matrix Φ as described in Sec. 4.1.

For each trial we quantize the measurements using a finite-range quantizer and use them to reconstruct the signal (a) by incorporating the saturated measurements in the reconstruction and (b) by discarding the saturated measurements before reconstruction. Both cases use the linear program (3) with the appropriate value for η. We denote the reconstructed signal with $\widehat{\mathbf{x}}_{\text{keep}}$ and $\widehat{\mathbf{x}}_{\text{discard}}$, respectively.

The results are shown in Fig. 1, which plots the average reconstruction SNR versus the quantizer dynamic range T for a variety of $\frac{M}{N}$. In particular, Figs. 1 (a) and (b) display the SNR of $\widehat{\mathbf{x}}_{\text{keep}}$ and $\widehat{\mathbf{x}}_{\text{discard}}$, respectively. Figure 1 (c) compares the two approaches for the two extreme cases of $\frac{M}{N} = \frac{3}{16}$ and $\frac{M}{N} = \frac{15}{16}$.

The plots demonstrate that lowering the threshold T such that the saturation rate is non-zero achieves a higher reconstruction SNR compared to scaling such that no measurements clip. Furthermore, rejecting saturated measurements performs better than incorporating them in the reconstruction. This is best illustrated in Fig. 1 (c): the optimal point on the dashed line, which corresponds to discarding saturated measurements, exhibits better SNR than the optimal point on the solid line, which corresponds to incorporating saturated measurements. As expected, the curves coincide when the saturation rate is effectively zero.

We also performed this experiment for larger values of K and B. As expected with higher B, we achieve less performance gain. As B grows, the quantization error goes down and thus reducing the quantization interval by dropping measurements is less effective. As K increases, rejecting measurements remains an optimal strategy. However, when K is large enough such that the non-saturated measurements do not satisfy RIP, our method performs worse than incorporating the saturated measurements.

5. Discussion

Our results demonstrate that CS overthrows the conventional wisdom on finite range quantization. Specifically the common practice of scaling the signal such that the ADC does not overflow is not optimal in light of the non-linear reconstruction. Our results demonstrate that allowing the signal to saturate is advantageous because it decreases the quantization interval in the unsaturated measurements. The non-linear reconstruction methods allow us to discard saturated measurements and prevent the saturation error from affecting the reconstruction process.
Our results further suggests a simple automatic gain control (AGC) strategy, in which the deviation of the average clipping rate from the desired one is used as a feedback to modify the gain. Since the desired clipping rate is non-zero, the feedback is symmetric and increases the gain if the clipping rate is too low. In comparison, classical AGC systems rely on the clipping rate only when the gain is too high and should be reduced. Since in such systems a zero clipping rate is the desired behavior, the AGC needs to rely on other signal features to ensure the gain is sufficient to provide a good signal-to-quantization noise ratio.

6. Acknowledgments

The work was supported by grants NSF CCF-0431150, CCF-0728867, CNS-0435425, and CNS-0520280, DARPA/ONR N66001-08-1-2065, ONR N00014-07-1-0936, N00014-08-1-1067, N00014-08-1-1112, and N00014-08-1-1066, AFOSR FA9550-07-1-0301, ARO MURI W311NF-07-1-0185, and the Texas Instruments Leadership University Program.

References:

[1] R. G. Baraniuk, M. A. Davenport, R. DeVore, and M. Wakin. A simple proof of the Restricted Isometry Property for random matrices. In *Constructive Approximation*, volume 28(3), pages 253–263, Dec 2008.

[2] E. Candes. Compressive sampling. In *Int. Congress of Mathematics*, volume 3, pages 1433–1452, 2006.

[3] S. Dasgupta and A. Gupta. An elementary proof of the Johnson-Lindenstrauss lemma. In *U.C. Berkeley Tech. Rep.*, volume TR-99-006, 1999.

[4] G. A. Gray and G. W. Zeoli. Quantization and saturation noise due to analog-to-digital conversion. In *IEEE Trans. on Aerospace and Electronic Systems*, pages 222–223, Jan 1971.

[5] J. N. Laska, P. Boufounos, M. A. Davenport, and R. G. Baraniuk. Democracy in action: finite-range scalar quantization for compressive sensing. In *To be submitted*, 2009.

[6] A. B. Sripad and D. L. Snyder. A necessary and sufficient condition for quantization errors to be uniform and white. In *IEEE Trans. on Acoustics, Speech, and Signal Processing*, volume ASSP-25, pages 442 – 448, 1977.

Special session on

Sampling
and
(In)Painting

Chair: Massimo FORNASIER

Report on Digital Image Processing for Art Historians

Bruno Cornelis [1,2], Ann Dooms [1,2], Ingrid Daubechies [2,3] and Peter Schelkens [1]

(1) Dept. of Electronics and Informatics (ETRO), Vrije Universiteit Brussel (VUB) - Interdisciplinary Institute for Broadband Technology (IBBT), Pleinlaan 2, B-1050 Brussels, Belgium.
(2) Computational and Applied Mathematics Program (CAMP), VUB
(3) Princeton University, Program in Applied and Computational Mathematics, Princeton, NJ 08544
bruno.cornelis@vub.ac.be

Abstract:

As art museums are digitizing their collections, a cross-disciplinary interaction between image analysts, mathematicians and art historians is emerging, putting to use recent advances made in the field of image processing (in acquisition as well as in analysis). An example of this is the *Digital Painting Analysis* (DPA) initiative [2], bringing together several research teams from universities and museums to tackle art related questions such as artist authentication, dating, etc. Some of these questions were formulated by art historians as *challenges* for the research teams. The results, mostly on van Gogh paintings, were presented at two workshops. As part of the Princeton team within the DPA initiative we give an overview of the work that was performed so far.

1. Introduction - Penetrating the art world

Determining the authenticity of a painting can be a daunting task for art historians, requiring extensive art historical research as well as the analysis of pigments, fabrics etc. However much insight chemical analysis yields [4], it requires the destruction of a sample from the painting and is therefore seldom allowed by conservators. Digital image processing for analyzing paintings could thus prove a useful addition to the art experts' toolbox, even beyond the purpose of authentication. We expect that art historians will gradually learn to use and trust these tools; a similar emergence and eventual success took place in the medical world in the mid 80s, with the advent of computed tomography. Subsequently, reconstruction algorithms played a significant role in creating other medical imaging technologies, including MRI, PET and SPECT.

To stimulate the interaction between the art historical world and branches of digital image processing, the *Digital Painting Analysis* (DPA) initiative organized two workshops in Amsterdam (IP4AI or *Image Processing for Artist Identification*) and a symposium (celebrating the inauguration of TiCC, *Tilburg centre for Creative Computing*) to facilitate a dialog between the two communities. The Van Gogh Museum (Amsterdam) and the Kröller Müller Museum (Otterlo) made it possible for participating teams to work with high resolution digital images of paintings (mostly van Goghs) in their collections.

2. Challenges - Convincing the art expert

To jumpstart the IP4AI workshops, art historians formulated *challenges* for the research teams, asking them to provide convincing arguments in favor of digital image processing. These included the following:

- *Authentication:* distinguish an original van Gogh painting from a copy or forgery. This was the main focus of the first workshop; preliminary results of the participating research teams can be found in [7].
- *Dating:* classify works by van Gogh that were either painted in his early Paris phase ($1886 - 1888$) or in his later Arles period. Art historians noticed changes in van Gogh's way of painting throughout his career. Small brushstrokes seem to be more prominent in his Paris period while broader ones prevail in Arles.
- *Identifying distinguishing features:* can an artist's hand be characterized and features be found that distinguish him from other painters?
- *Image enhancement:* fuse information obtained by different modalities (x-ray, infrared, visual, etc.) to (virtually) enhance damaged paintings, or underpaintings. A first challenge here is detailed and precise registration.
- *Inpainting:* digitally reconstruct missing pieces from a painting when only limited data is at hand.

The purpose of this paper is to provide an overview of the tools and general methodology used by the Princeton research team in order to tackle these challenges. Detailed results can be found in [6, 10].

2.1 Classification - Authentication, dating and identifying features

For the analysis of paintings it is crucial to extract distinguishing features/statistics that truly characterize the style of an artist. It is obvious that simple image statistics such as mean or variance of an image will not suffice by themselves. To take an extreme example: reordering by increasing grayscale the pixels in every row of a digital image of a natural scene, and then doing the same in every column, produces an image with same mean and variance as the original, but bereft of (almost) all other information. More complex models that provide additional information

are needed. The approach taken for the first three challenges built such models. The analysis consists of *three main steps*: transform, modeling and classification.

Transform. A multiresolution transform is performed on patches of the image. We used the Dual-Tree Complex Wavelet Transform (DTCWT) [9]; it provides approximate shift invariance and directional selectivity (properties standard wavelet transforms lack). The DTCWT uses two parallel filter banks and produces six subbands of coefficients that let us analyze changes in the image in six directions ($\pm 15°, \pm 45°$and $\pm\ 75°$) at different scales.

Modeling. A large number of pixels, and thus also of transform coefficients, combined with noise on the pixel values (due to the acquisition process) impose robust dimensionality reduction and feature extraction techniques. We used Hidden Markov Trees (HMT) [3]. It is possible to describe the wavelet coefficients for a large class of images in terms of two key properties [11]:

- *2Population:* smooth image regions are represented by wavelet coefficients with a narrow probability distribution function (pdf); edges, ridges or other singularities by wavelet coefficients with a wide pdf.
- *Persistence:* the classification into narrow/wide pdf-coefficients tends to propagate across scales.

These two properties are used to design a statistical model to represent images. Due to the multiresolution nature of the wavelet transform, the wavelet coefficients can be arranged into a quadtree (one coefficient from a coarser scale corresponds to four wavelet coefficients at the next finer scale). At each scale, hidden variables control the wavelet coefficients. They can have two states: L (large, for edge-like structures) and S (small, for smooth regions). The wavelet coefficients are modeled as samples from a mixture of two Gaussian distributions, one with a large variance for the coefficients corresponding to an edge and one with a small variance for coefficients from a smooth region. HMT model the statistical dependencies between wavelet coefficients at different scales. The parameters of the HMT we used as features are:

- α_T: a 2×2 transition probability matrix, that depicts the probabilities that a child node is in a particular state, given the state of the parent node.
- μ_i: the means of the narrow and wide Gaussian distribution ($i = 1, 2$) for each subband.
- σ_i: variance of the narrow and wide Gaussian distribution ($i = 1, 2$) for each subband.

For example, if we apply a 4-level DTCWT transform on a patch of an image, then the features extracted from that patch would be the following: $6 \times 4 \times 2$ means, $6 \times 4 \times 2$ variances and $6 \times 4 \times (2 \times 2)$ probabilities, adding up to a total of 192 features. These HMT features are grouped into a model parameter vector and are determined using the *expectation maximization* algorithm.

Classification. The model parameters vectors extracted in the previous step are used as the input for classification algorithms. We used several types of machine learning algorithms: Support Vector Machines, Adaboost, Decision Stump and Random Forest. All experiments were performed with WEKA [1], a collection of machine learning algorithms for data mining tasks.

2.1.1 Authentication challenge results

The authentication challenge was the main research topic for the first IP4AI workshop [7]. To validate their earlier results the Princeton team asked Dutch art conservation student Charlotte Caspers to make original paintings on different materials, with different kinds of paint and brushes, and to create a faithful copy for each of these originals. The dataset provided ground truth: we knew which paintings were original and which ones were copies. We considered both HMT features and *thresholding features* [10]. The aim was to recognize the difference between a fluid and a more hesitant (copying) stroke through machine learning. For this kind of classification problem the *SVM with polynomial kernel* machine learning algorithm was the best classifier.
The images were subdivided into patches, some of which were used for training the machine learning algorithm. The best results were obtained by using only patches from the painting under investigation and its copy (see Figure 1). The results can be found in Table 1; they show that when both soft and hard brushes are used, the algorithm achieves a succes rate similar to that obtained by state-of-the-art authentication algorithms for handwriting.

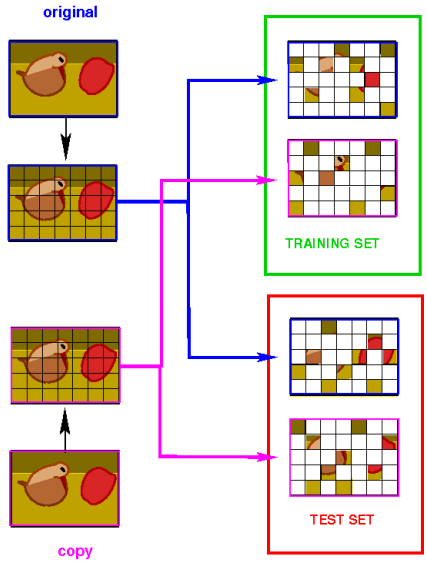


Figure 1: Four sets of patches without overlap.

2.1.2 Dating challenge results

For the dating challenge a set of 66 high resolution paintings (90 pixels per linear inch) were put at the disposal of all teams. All the classifiers listed above were trained with 256×256 patches using 10-fold cross validation. As can be seen in Table 2, the Random Forest (RF) classifier was the most accurate. Three paintings for which art historians are not sure when they were painted needed to be attributed to one of two periods. Figure 2 shows the resulting classification success rate for patches of paintings from the training set. The RF algorithm was then used on the patches of the three paintings to be attributed, and a majority vote of the patches was determined.

Pair	Ground	Paint	Brushes	Style	Total	Copy	Original
1	CP Canvas	Oils	Soft&Hard		78%	67%	89%
2	CP Canvas	Acrylics	Soft&Hard		72%	55%	89%
3	Smooth CP Board	Oils	Soft&Hard		78%	78%	78%
4	Bare linen canvas	Oils	Soft	TI	75%	50%	100%
5	Chalk and Glue	Oils	Soft	TI	50%	0%	100%
6	CP Canvas	Acrylics	Soft	TI	38%	75%	0%
7	Smooth CP Board	Oils	Soft	Sm,Bl	55%	22%	88%

Table 1: Accuracy for each test on the Caspers data set. Abbreviations: Sm=Smooth, Bl=Blended, TI=Thick Impasto.

SVM	AB	DS	RF
61.2%	63.2%	63.1%	70.5%

Table 2: Accuracy of different classifiers. Abbreviations: SVM=Support Vector Machines, AB= AdaBoost, DS=Decision Stump, RF=Random Forest.

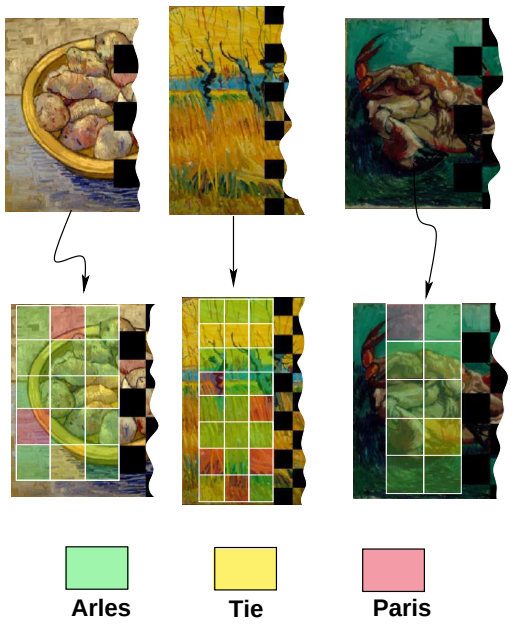


Figure 2: Classification results for three paintings.

2.1.3 Extracting Distinguishing Features results

The test set consisted of floral still lifes painted by van Gogh, Monticelli and other contemporary artists. The goal was to quantify to what extent van Gogh and Monticelli share features, in their brushwork and color schemes, absent in the style of the others. The purpose here was thus to distinguish styles instead of painters (as in authentication). The same methodology described above was used. Results show that wavelet coefficients in direction $-45°$, scale 6 characterize the style of van Gogh and Monticelli whereas wavelet coefficients in the $15°$, scale 4 subband are more prominent in the other paintings. Examples of these distinguishing features are highlighted in Figure 3. More detailed results are in [6].

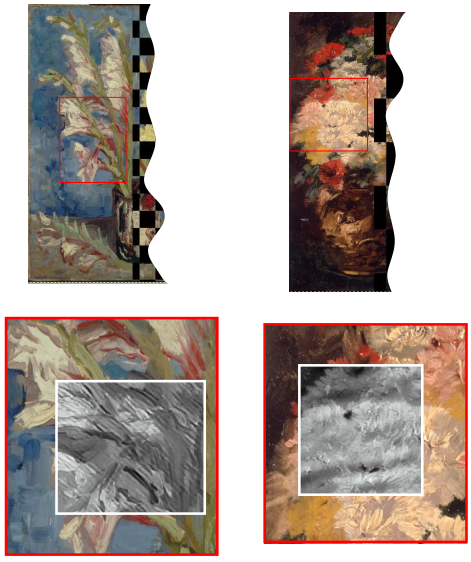


Figure 3: Distinguishing feature challenge. Left: "Still Life: Vase with Gladioli" by V. van Gogh. Right: "Vase with Flowers" by G. Jeannin.

2.2 Using Different Image Acquisitions

Art museums typically have x-ray and infrared photographs in their collections, which can reveal much about what is below the visible surface of a painting. These can also be digitized (or acquired digitally, in the future), and be studied with digital image processing tools. In order to combine the different modes of image acquisition, the first task is to register the images (we used x-ray, infrared and color images of the same painting) to enhance and detect hidden features. Figure 4 shows a woman's face emerging (horizontally) from underneath the grass in the painting "Patch of Grass". Because x-rays and photographs are acquired by different modalities, the matching is not as straightforward as it seems initially. Both images were divided into patches and reference points in both images were picked in order to define a smooth warping that gave acceptable results. Another example is the counting of threads/inch in the canvas, visible on x-rays, to determine a painting's authenticity and date [8].

2.3 Inpainting

An important aspect for art historians and conservators is the preservation of works of art. When paintings become damaged, all the available information (grayscale photographs, low resolution color photographs, ektachromes,

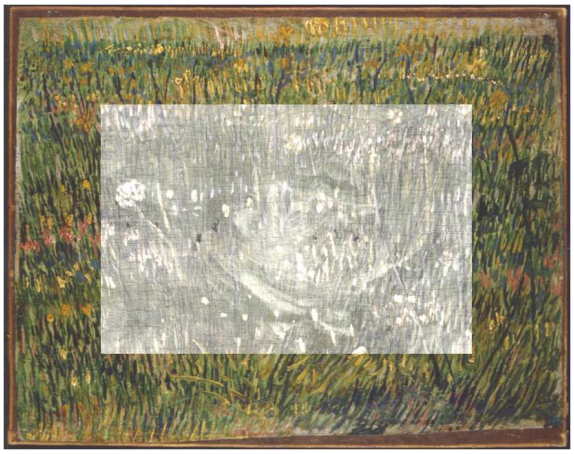


Figure 4: Registered x-ray on "Patch of Grass".

...) is called upon to help art conservators in their reconstruction or restoration. In [5] techniques were proposed to *mathematically* reconstruct the original colors of frescoes (reduced to rubble in a wartime bombing) by making use of the information given by preserved fresco fragments and gray level pictures of the intact frescoes taken before the damage occurred.

We investigated whether such techniques would also work on van Gogh pictures. With the help of M. Fornasier, one of the authors of [5], we applied these algorithms to a high resolution color image of the "Lemons on a Plate" painting. A patch of 200×200 pixels was digitally removed; Figure 5 shows its mathematical reconstruction, using only a low resolution color image (with faithful colors) and a high resolution grayscale image of that painting. The results are quite satisfying and prove that these techniques could be used for restoration purposes.

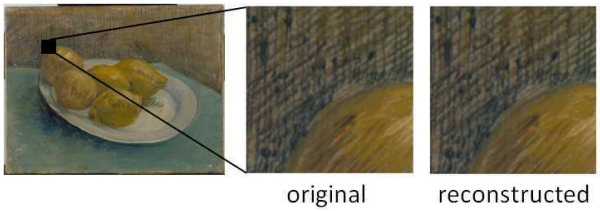


Figure 5: Inpainting.

3. Conclusions

The results obtained for the first and second IP4AI workshop in Amsterdam were promising. It is clear however, that these digital techniques on their own are not sufficient to provide conclusive answers to questions of interest to art historians. Nevertheless, they will likely be a worthy addition to the toolbox of art historians and conservators; they have the great advantage of not being invasive. There is also still room for improvement in the different steps of the analysis of paintings. It is worth pointing out, however, that in order to apply such techniques, the quality of the acquired dataset (i.e. high resolution images) is of utmost importance. Only images of equal quality can be compared with each other.

4. Acknowledgments

We would like to thank Sina Jafarpour, Gungor Polatkan, Andrei Brasoveanu, Eugene Brevdo and Shannon M. Hughes for letting us report briefly on some of their results, and the Van Gogh and Kröller Müller Museums for letting us use their high resolution data set. Special thanks go to Massimo Fornasier for his help with the inpainting challenge.

Research was partially supported by The Fund for Scientific Research Flanders (project G.0206.08 and postdoctoral fellowship of Peter Schelkens).

References:

[1] http://www.cs.waikato.ac.nz/ml/weka/.

[2] http://www.digitalpaintinganalysis.org/.

[3] Matthew Crouse, Robert Nowak, and Richard Baraniuk. Wavelet-based statistical signal processing using hidden markov models. *IEEE Transactions on Signal Processing*, 46:886–902, 1997.

[4] Joris Dik, Koen Janssens, Geert Van Der Snickt, Luuk van der Loeff, Karen Rickers, and Marine Cotte. Visualization of a lost painting by vincent van gogh using synchrotron radiation based x-ray fluorescence elemental mapping. *Anal. Chem.*, 80:6436–6442, 2008.

[5] Massimo Fornasier, Ronny Ramlau, and Gerd Teschke. A comparison of joint sparsity and total variation minimization algorithms in a real-life art restoration problem. *to appear in Advances in Computational Mathematics*, 2008.

[6] S. Jafarpour, G. Polatkan, E. Brevdo, S. Hughes, A. Brasoveanu, and I. Daubechies. Stylistic analysis of paintings using wavelets and machine learning. *submitted to EUSIPCO 2009*.

[7] C. R. Johnson, Ella Hendriks, Igor Berezhnoy, Eugene Brevdo, Shannon Hughes, Ingrid Daubechies, Jia Li, Eric Postma, and James Z. Wang. Image processing for artist identification - computerized analysis of vincent van gogh's painting brushstrokes. *IEEE Signal Processing Magazine*, July 2008.

[8] D. H. Johnson, L. Sun, J. Guo, C. R. Johnson Jr., and E. Hendriks. Matching canvas weave patterns from processing x-ray images of master paintings. *submitted to 16th IEEE International Conf. on Image Processing*, 25:37–48, November 2009.

[9] Nick Kingsbury. Complex wavelets for shift invariant analysis and filtering of signals. *Applied and Computational Harmonic Analysis*, 10(3):234–253, May 2001.

[10] G. Polatkan, S. Jafarpour, A. Brasoveanu, S. Hughes, and I. Daubechies. Detection of forgery in paintings using supervised learning. *Submitted to IEEE International Conference on Image Processing 2009*.

[11] Justin K. Romberg, Hyeokho Choi, Richard G. Baraniuk, and Nick Kingsbury. A hidden markov tree model for the complex wavelet transform. *IEEE Transactions on Signal Processing*, pages 133–136, 2001.

Edge Orientation Using Contour Stencils

Pascal Getreuer [1]

(1) Department of Mathematics, University of California Los Angeles
getreuer@math.ucla.edu

Abstract:

Many image processing applications require estimating the orientation of the image edges. This estimation is often done with a finite difference approximation of the orthogonal gradient. As an alternative, we apply contour stencils, a method for detecting contours from total variation along curves, and show it more robustly estimates the edge orientations than several finite difference approximations. Contour stencils are demonstrated in image enhancement and zooming applications.

1. Introduction

A fundamental and challenging problem in image processing is estimating edge orientations. Accurate edge orientations are important for example in edge-oriented inpainting methods [2], and optical character recognition features [8].

1.1 $\nabla u^\perp$ for Estimating Edge Orientation

A starting point to edge orientation estimation is to approximate $\nabla u^\perp$ with finite differences. Finite difference estimation alone is typically too noisy to be reliable, especially near edges, so the gradient is often regularized by a convolution $\nabla u \approx \nabla(G * u)$ where G is for example a Gaussian. However, there is a serious problem in that $\nabla u^\perp$ and $-\nabla u^\perp$ both describe the same edge orientation, so linear smoothing tends to *cancel* the desired edge information.

Introduced by Bigün and Granlund [1] and Forstner and Gulch [3], a better approach is to use the 2×2 *structure tensor* $J(\nabla u) = \nabla u \otimes \nabla u$. The structure tensor satisfies $J(-\nabla u) = J(\nabla u)$ and ∇u is an eigenvector of $J(\nabla u)$. The structure tensor takes into account the orientation but not the sign of the direction, thus solving the antipodal cancellation problem.

As developed by Weickert [9], let

$$J_\rho(\nabla u_\sigma) = G_\rho * J(G_\sigma * u) \tag{1}$$

where G_σ and G_ρ are Gaussians with standard deviations σ and ρ. The eigenvector of $J_\rho(\nabla u_\sigma)$ associated with the smaller eigenvalue is called the *coherence direction*, and is an effective approximation of edge orientation.

2. Contour Stencils

Numerical implementation of $J(\nabla u)$ yet involves estimating ∇u. Since numerical estimates of ∇u are sensitive to noise and unreliable near edges, significant amounts of smoothing is still needed for acceptable results. We abandon $\nabla u^\perp$ and approach the estimation of edge orientation from an entirely different principle.

Given a smooth curve C and a parameterization $\gamma : [0, T] \to C$, consider measuring the total variation of u along C,

$$\mathrm{TV}(C) = \int_0^T \left|\partial_t u\big(\gamma(t)\big)\right| \, dt. \tag{2}$$

Edge orientations can be estimated by comparing $\mathrm{TV}(C)$ with various candidate curves. *Contour stencils* [4, 5] is a numerical implementation of this idea.

Let $u : \Lambda \to \mathbb{R}$ be a discrete image. Denote by $u_{i,j}$, $(i, j) \in \Lambda$, the value of u at the (i, j)th pixel, and let $x_{i,j} \in \mathbb{R}^2$ denote its spatial location.

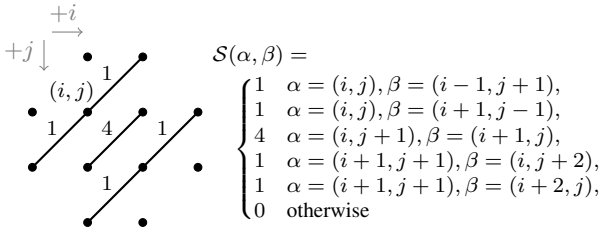


Figure 1: An example contour stencil $\mathcal{S}$ for detecting a $45°$ orientation.

A *contour stencil* is a function $\mathcal{S} : \Lambda \times \Lambda \to \mathbb{R}^+$ describing weighted edges between pixels (see Figure 1). These edges approximate several parallel curves localized over a small neighborhood. As a discretization of (2), the total variation of $\mathcal{S}$ is

$$\mathrm{TV}(\mathcal{S}) := \tfrac{1}{|\mathcal{S}|} \sum_{\alpha,\beta\in\Lambda} \mathcal{S}(\alpha,\beta)\,|u_\alpha - u_\beta|\,, \qquad (3)$$

and $|\mathcal{S}| := \sum_{\alpha,\beta} \mathcal{S}(\alpha,\beta)\,|x_\alpha - x_\beta|$. For the contour stencil in Figure 1, $|\mathcal{S}| = (1+1+4+1+1)\sqrt{2}$ and

$$\begin{aligned}\mathrm{TV}(\mathcal{S}) = \tfrac{1}{|\mathcal{S}|}\big(&|u_{i,j} - u_{i-1,j+1}| + |u_{i,j} - u_{i+1,j-1}| \\ &+ 4\,|u_{i,j+1} - u_{i+1,j}| \\ &+ |u_{i+1,j+1} - u_{i,j+2}| + |u_{i+1,j+1} - u_{i+2,j}|\big).\end{aligned}$$

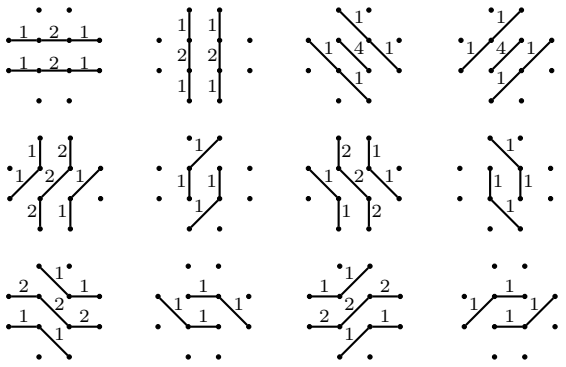


Figure 2: The proposed cell-centered contour stencils.

The contours of u are estimated by finding a stencil with low total variation,

$$\mathcal{S}^* = \arg\min_{\mathcal{S}\in\Sigma} \mathrm{TV}(\mathcal{S}) \qquad (4)$$

where Σ is a set of candidate stencils (see Figures 2 and 3). The best-fitting stencil $\mathcal{S}^*$ provides a model of the underlying contours.

In summary, contour stencil orientation estimation is done by first computing the TV estimates (3) for each candidate stencil, and then determining the best-fitting stencil $\mathcal{S}^*$. For efficient implementation, define

$$\begin{aligned} D^H_{i,j} &= |v_{i,j} - v_{i+1,j}|\,, & D^A_{i,j} &= |v_{i,j} - v_{i+1,j+1}|\,, \\ D^V_{i,j} &= |v_{i,j} - v_{i,j+1}|\,, & D^B_{i,j} &= |v_{i,j+1} - v_{i+1,j}|\,, \end{aligned}$$

then the $\mathrm{TV}(\mathcal{S})$ can be computed as sums of these differences, and the differences may be reused between successive cells. For the proposed stencil sets, contour stencils cost a few dozen operations per pixel [4].

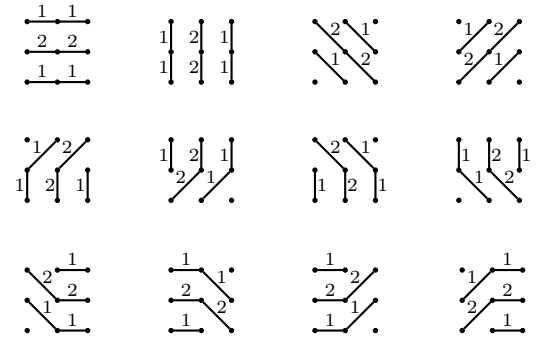


Figure 3: A node-centered stencil set.

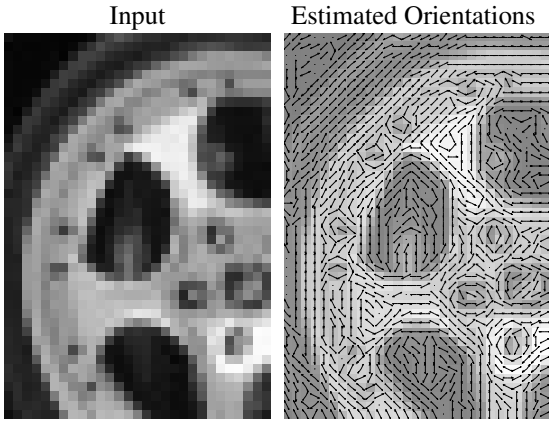


Figure 4: Edge orientation estimation with contour stencils (using the cell-centered stencils in Figure 2).

Contour stencils extend naturally to nonscalar data by replacing the absolute value in (3) with a metric. On color images for example, a suitable choice is the ℓ^1 vector norm in $\mathrm{YC_bC_r}$ color space.

3. Comparison

Here we compare contour stencils and several finite difference methods for estimating edge orientation.

As a test image with fine orientations, we use a small image of straw (Figure 5).

u

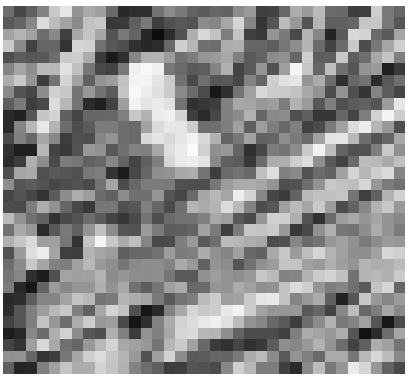
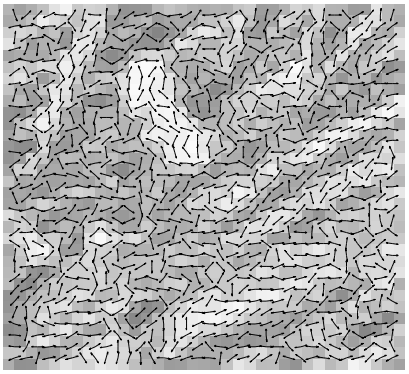


Figure 5: The test image.

As is done with coherence direction (1), any orientation field $\vec{\theta}$ can be smoothed by filtering its tensor product: $G_\rho * (\vec{\theta} \times \vec{\theta})$. But for easier comparison, all methods are shown *without* smoothing.

$\nabla u^\perp$ with (5)



Contour Stencils (Σ as in Figure 2)

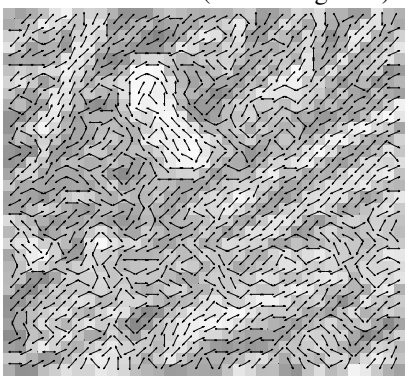


Figure 6: Comparison of cell-centered methods.

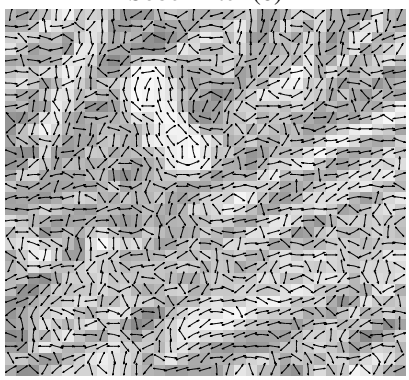
We consider two categories of methods: cell-centered and node-centered. Define the (i,j)th cell as the square whose corners correspond to $u_{i,j}$, $u_{i+1,j}$, $u_{i,j+1}$, $u_{i+1,j+1}$. Cell-centered methods compute orientation estimates logically located in the center of the cells. With node-centered methods, the edge orientation estimates are centered on the pixels.

Let D_x^+ denote the forward difference operator $D_x^+ u_{i,j} = u_{i+1,j} - u_{i,j}$ and similarly in the other coordinate D_y^+. An estimate of ∇u symmetric over the cell is

$$\nabla u_{i,j} \approx \begin{pmatrix} (D_x^+ u_{i,j} + D_x^+ u_{i,j+1})/2 \\ (D_y^+ u_{i,j} + D_y^+ u_{i+1,j})/2 \end{pmatrix}. \qquad (5)$$

Figure 6 compares $\nabla u^\perp$ estimated using (5) with contour stencils using the cell-centered stencil set shown in Figure 2.

Sobel filter (6)



Contour Stencils (Σ as in Figure 3)

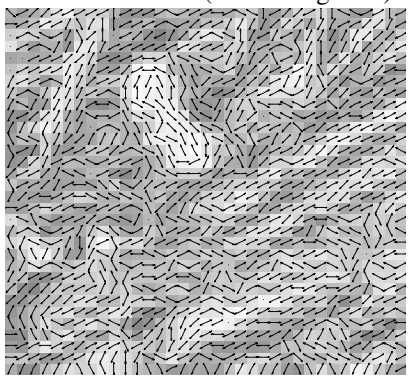


Figure 7: Comparison of node-centered methods.

The Sobel filter [7] is a node-centered approximation of ∇u,

$$\partial_x u \approx \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} * u \qquad (6)$$

and similarly for $\partial_y u$. Figure 7 compares the Sobel filter with contour stencils using the node-centered stencil set from Figure 3.

4. Applications

Contour stencils are useful in applications where edges are significant.

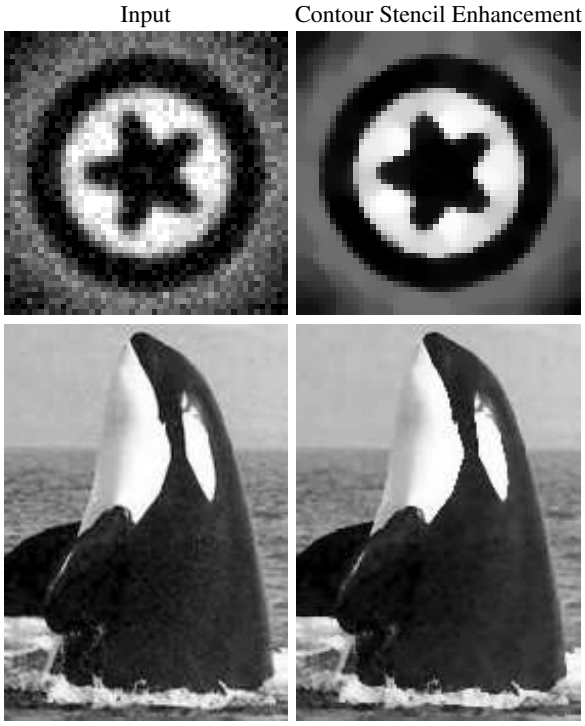


Figure 8: Simultaneous sharpening and denoising using contour stencils [4].

Contour stencils can be useful in discretizing image diffusion processes. Figure 8 demonstrates image enhancement using a combination of the Rudin-Osher shock filter [6] and TV-flow that has been discretized with contour stencils.

As another application, Figure 9 shows an image zooming result using contour stencils. The method approaches zooming as an inverse problem using a least-squares graph regularization. The regularization is adapted according to the edge orientations estimated from the contour stencils.

5. Conclusions

Contour stencils provide reliable orientation estimates at low computational cost, enabling better results in image processing applications.

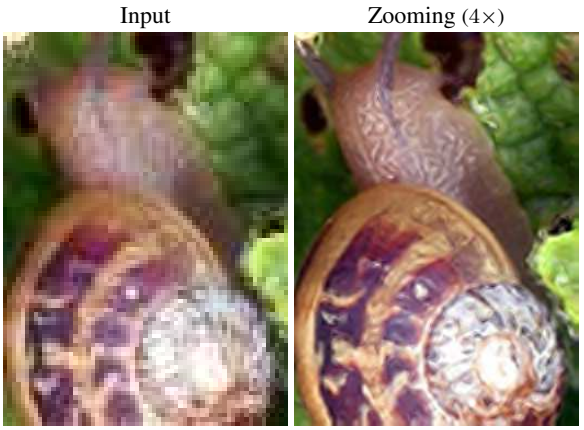


Figure 9: (This is a color image.) Edge-adaptive zooming using contour stencils [5].

References:

[1] J. Bigün and G. H. Granlund. Optimal orientation detection of linear symmetry. In *IEEE First International Conference on Computer Vision*, pages 433–438, London, Great Britain, June 1987.

[2] Folkmar Bornemann and Tom März. Fast image inpainting based on coherence transport. *J. Math. Imaging Vis.*, 28(3):259–278, 2007.

[3] W. Förstner and E. Gulch. A fast operator for detection and precise location of distinct points, corners, and centers of circular features. pages 281–305, 1987.

[4] Pascal Getreuer. Contour stencils for edge-adaptive image interpolation. volume 7257, 2009.

[5] Pascal Getreuer. Image zooming with contour stencils. volume 7246, 2009.

[6] S. J. Osher and L. I. Rudin. Feature-oriented image enhancement using shock filters. *SIAM Journal on Numerical Analysis*, 27:919–940, 1990.

[7] Irwin Sobel and Jerome A. Feldman. A 3x3 isotropic gradient operator for image processing. Presented at a talk at the Stanford Artificial Project in 1968.

[8] Øivind Due Trier, Anil K. Jain, and Torfinn Taxt. Feature-extraction methods for character-recognition: A survey. *Pattern Recognition*, 29(4):641–662, April 1996.

[9] Joachim Weickert. *Anisotropic diffusion in image processing*. ECMI Series, Teubner-Verlag, Stuttgart, Germany, 1998.

Smoothing techniques for convex problems. Applications in image processing.

Pierre Weiss [(1)], Mikaël Carlavan [(2)], Laure Blanc-Féraud [(2)] and Josiane Zerubia [(2)]

(1) Institute for Computational Mathematics, Kowloon Tong, Hong Kong.
(2) Projet Ariana - CNRS/INRIA/UNSA, 2004 route des Lucioles, 06902 Sophia-Antipolis, France.
(1) pierre.armand.weiss@gmail.com, (2) firstname.lastname@sophia.inria.fr

Abstract:

In this paper, we present two algorithms to solve some inverse problems coming from the field of image processing. The problems we study are convex and can be expressed simply as sums of l^p-norms ($p \in \{1, 2, \infty\}$) of affine transforms of the image. We propose 2 different techniques. They are - to the best of our knowledge - new in the domain of image processing and one of them is new in the domain of mathematical programming. Both methods converge to the set of minimizers. Additionally, we show that they converge at least as $O\left(\frac{1}{N}\right)$ (where N is the iteration counter) which is in some sense an "optimal" rate of convergence. Finally, we compare these approaches to some others on a toy problem of image super-resolution with impulse noise.

1. Introduction

Many image processing tasks like reconstruction or segmentation can be done efficiently by solving convex optimization problems. Recently these models received considerable attention and this led to some breakthrough. Among them are the new sampling theorems [5] and the impressive results obtained using sparsity or regularity assumptions in image reconstruction (see e.g. [4]).

These results motivate an important research to accelerate the convergence speed of the minimization schemes. In the last decade, many algorithms like iterative thresholding or dual approaches were reinvented by the "imaging community" (see for instance [2, 3] for old references). Recently, the "mathematical programming community" got interested in those problems and it led to some drastic improvements. As examples let us cite the papers by Y. Nesterov [9, 10] and M. Teboulle [1] which improve by one order of magnitude most first order approaches.

In this paper, we mainly follow the lines of Y. Nesterov [9]. We consider the problem of minimizing the sum of l^p-norms ($p \in \{1, 2, \infty\}$) of affine transforms of the image. The general mechanism of the algorithms we propose consists in smoothing the problem and solve it with an efficient first order scheme. Our contribution is mainly to extend the results of [9] to a more general setting and to propose a dual variant which behaves better in all problems we tested. We also give convergence rates for the proposed algorithms. We believe, this gives some insight on the important factors that influence the algorithms efficiency and helps designing solvable problems.

2. The problems considered

In this paper, we consider the following seminal model of image deterioration:

$$u^0 = Du + b \tag{1}$$

where u is an original, neat image, $D : \mathbb{R}^n \to \mathbb{R}^m$ is some known linear transform, $b \in \mathbb{R}^m$ is some additive noise and $u^0 \in \mathbb{R}^m$ is a given observed image. This simple formalism actually models many real situations. For instance, D can be an irregular sampling and a convolution. In this case recovering u from u^0 is a super-resolution problem [7]. Other applications include image inpainting, compression noise reduction, texture+cartoon decompositions, reconstruction from noisy indirect measurements... Finding u from the observation u^0 is an inverse problem. There exists many ways to solve it. In this paper, we concentrate on two variational models. The first one consists in solving the following convex problem:

$$\min_{x \in X} \left(\underbrace{||Bx||_1 + \lambda ||Dx - u^0||_p}_{\Psi(x)} \right). \tag{2}$$

The second one consists in solving:

$$\min_{y \in Y} \left(||y||_1 + \lambda ||DB^* y - u^0||_p \right). \tag{3}$$

In both problems, $B : \mathbb{R}^n \to \mathbb{R}^o$ is a linear transform, $|| \cdot ||_p$ denotes the standard l^p-norm and X and Y are simple convex sets (like $\mathbb{R}^n$ or $[0, 1]^n$).

The interpretation of the first model is as follows: we look for an image x which minimizes $||Bx||_1$ such that Dx is close to u^0. The function $x \mapsto ||Bx||_1$ can be seen as a *regularity* a priori on the image. For instance, if B is the discrete gradient, then it corresponds to the total variation. If B is some wavelet transform, it is equivalent to a Besov semi-norm [6]. p must be chosen depending on the statistics of the additive noise. For instance, p should be equal to 2 for Gaussian noise, to 1 for impulse noise and to ∞ for uniform noise.

The interpretation of the second model is the following: we look for a decomposition y of the restored image in some dictionary B^* such that its reconstruction $B^* y$ is close to u^0. Minimizing the l^1-norm of y is known to favor sparse structures. The underlying assumption is thus that the original image u is sparse in the dictionary B^*.

From a numerical point of view, both problems are very similar. However, the first one is slightly more general and complicated than the second. We will thus give a detailled analysis of its resolution and only provide numerical results for the second one.
The remaining of the paper is as follows. We first present an algorithm based on a regularization of the primal problem (2). Then we present a technique to regularize a dual version of (2). Finally we propose theoretical and numerical comparisons of both techniques on a problem of image super-resolution. Due to space limitations, we only provide the main ideas in this paper. We refer the reader to [12] (in French), for the proofs of the propositions.

3. Smoothing of the primal problem

In this section, we propose a method to minimize (2). Its principle is exactly the same as the method proposed by Y. Nesterov in [9]:

1. Smooth the non-differentiable terms in (2).
2. Solve the regularized problem using an accelerated gradient method.

The only difference is that we do not require the set X to be bounded, which requires a slightly different analysis. Now let us present some details of this approach. A key observation to solve (2) is that it can be rewritten as a so called min-max problem. Let p' denote the conjugate of p $\left(\text{i.e. } \frac{1}{p'}+\frac{1}{p}=1\right)$. We can rewrite problem (2) as follows:

$$\min_{x\in X}\left(\max_{y\in Y}\left(\langle Bx,y_1\rangle+\lambda\langle Dx-u^0,y_2\rangle\right)\right) \quad (4)$$

$$= \min_{x\in X}\left(\underbrace{\max_{y\in Y}\left(\langle Ax-h,y\rangle\right)}_{\Psi(x)}\right) \quad (5)$$

where $\langle\cdot,\cdot\rangle$ denotes the canonical scalar product,

$$A=\begin{bmatrix} B \\ \lambda D\end{bmatrix},\quad h=\begin{bmatrix} 0 \\ \lambda u^0\end{bmatrix} \text{ and} \quad (6)$$

$$Y=\{y=(y_1,y_2)\in\mathbb{R}^o\times\mathbb{R}^m, ||y_1||_\infty\leq 1, ||y_2||_{p'}\leq 1\}. \quad (7)$$

The function Ψ is a conjugate function and the set Y is bounded. It can thus be smoothed using a Moreau regularization. Let us denote:

$$\Psi_\mu(x)=\max_{y\in Y}\left(\langle Ax-h,y\rangle-\frac{\mu}{2}||y||_2^2\right). \quad (8)$$

This function can be shown to be L-Lipschitz differentiable:

$$||\nabla\Psi_\mu(x_1)-\nabla\Psi_\mu(x_2)||_2\leq L||x_1-x_2||_2 \quad (9)$$

with $L=\frac{|||A|||^2}{\mu}$ and $|||A|||=\max_{x\in\mathbb{R}^n,||x||_2\leq 1}(||Ax||_2)$. Furthermore, it is a good uniform approximation of Ψ in the following sense:

$$0\leq\Psi(x)-\Psi_\mu(x)\leq\frac{\mu}{2}D. \quad (10)$$

where $D=\left(\max_{y\in Y}\left(||y||_2^2\right)\right)$. Thus, we can make the difference between Ψ and Ψ_μ as small as desired by decreasing μ. The approximation Ψ_μ is actually very common in image processing. For instance, when $p=1$, it corresponds to the approximation of the absolute value by a Huber function. When $p=\infty$ it is slightly more difficult, but it can still be computed in closed form.
The smoothed problem writes:

$$\min_{x\in X}\left(\Psi_\mu(x)\right). \quad (11)$$

It consists in minimizing a differentiable function over a simple set. We can thus apply projected gradient like algorithms to solve it. Unfortunately, μ has to be chosen small in order to get a good approximate solution. This requires to use small step sizes in the gradient descent and thus results in a very slow convergence rate. The main observation of Y. Nesterov in [9] is that using an accelerated version of the projected gradient methods can actually compensate the approximation error. This results in a convergence rate in $O\left(\frac{1}{N}\right)$ (where N is the iteration counter), while other first order approaches like projected subgradient descents converge as $O\left(\frac{1}{\sqrt{N}}\right)$.
Now let us write down the complete algorithm to solve (11). Let x_μ^* denote a solution of (11) (it is not unique in general). We propose the following algorithm:

Algorithm 1 (Primal)

Choose a number of iterations N.
Set a starting point x^0 (as close as possible to x_μ^*).
Set $\mu=\mu(N)=\frac{|||A|||\cdot||x^0-x_\mu^*||_2}{N}$.
Set $\mathcal{A}=0$, $g=0$ and $x=x^0$.
for $k=0$ to N **do**
  $a=\frac{1}{L}+\sqrt{\frac{1}{L^2}+\frac{2}{L}\mathcal{A}}$
  $v=\Pi_X\left(x^0-g\right)$
  $y=\frac{\mathcal{A}x+av}{\mathcal{A}+a}$
  $x=\Pi_X\left(y-\frac{\nabla\Psi_\mu(y)}{L}\right)$
  $g=g+a\nabla\Psi_\mu(x)$
  $\mathcal{A}=\mathcal{A}+a$
end for
Set $x^N=x$.

Our main convergence results are as follows. Let x^* denote a solution of (2).

Proposition 1 *x^N converges to the set of minimizers of (2).*

Proposition 2 *The worst case convergence rate is:*

$$\Psi(x^N)-\Psi(x^*)\leq\frac{2|||A|||\cdot||x^0-x_\mu^*||_2\sqrt{D}}{N}. \quad (12)$$

Note that the distance $||x^0-x_\mu^*||_2$ is unknown in general, so that Algorithm 1 might not seem implementable. In the case where X is a compact set, this quantity can be bounded above by the diameter of X. When X is not bounded, it actually suffices to choose μ of order $\frac{|||A|||}{N}$ to

get a precision of order $O\left(\frac{1}{N}\right)$. Algorithm (1) is thus implementable and converges as $O\left(\frac{1}{N}\right)$. This convergence rate is neatly sublinear and might seem bad at first sight. Actually, it is somehow optimal. Indeed, A. Nemirovski shows in [8] that some instances of problems like (5) cannot be solved with a better rate of convergence than $O\left(\frac{1}{N}\right)$ using first order methods.

4. Smoothing of the dual problem

In this section, we propose an approach alternative to the previous one. Its flavor is similar to a proximal-method. One way to understand this scheme is that we smooth the "dual" problem instead of the primal problem. Note that the $\min$ and the $\max$ in equation (5) cannot be inverted as we do not suppose X to be compact. So we cannot use - properly speaking - the term dual problem.
Instead of solving (2), we solve:

$$\min_{x\in X}\left(||Bx||_1+\lambda||Dx-u^0||_p+\frac{\epsilon}{2}||x-x^0||_2^2\right) \quad (13)$$

where $\epsilon\in\mathbb{R}_*^+$ and x^0 should be chosen close to the set of minimizers of (2). It can be shown that as ϵ goes to 0, the unique solution of (13) converges to the Euclidean projection of x^0 onto the set of minimizers of (2). We can rewrite (13) as a min-max problem:

$$\min_{x\in X}\left(\max_{y\in Y}\left(\langle Ax-h,y\rangle\right)+\frac{\epsilon}{2}||x-x^0||_2^2\right) \quad (14)$$

$$= \max_{y\in Y}\left(\underbrace{\min_{x\in X}\left(\langle Ax-h,y\rangle+\frac{\epsilon}{2}||x-x^0||_2^2\right)}_{\Psi_\epsilon(y)}\right) \quad (15)$$

Note that we can invert the $\min$ and the $\max$ only because the term $\frac{\epsilon}{2}||x-x^0||_2^2$ makes the problem coercive in x. Now, the important observation is that the function Ψ_ϵ is the conjugate of a strongly convex function. It is thus concave and Lipschitz differentiable:

$$||\nabla\Psi_\epsilon(y_1)-\nabla\Psi_\epsilon(y_2)||_2\leq L||x_1-x_2||_2 \quad (16)$$

$\forall(y_1,y_2)\in Y\times Y$ with $L\leq\frac{|||A|||^2}{\epsilon}$. Problem (15) consists in maximizing a Lipschitz differentiable concave function over a convex set. It thus seems interesting to use a scheme similar to Algorithm 1 on this problem. Unfortunately we will get a convergence rate on the dual variable y and not on the variable of interest:

$$x(y)=\arg\min_{x\in X}\left(\langle Ax-h,y\rangle+\frac{\epsilon}{2}||x-x^0||_2^2\right). \quad (17)$$

Actually, a slight modification of Nesterov's scheme (an ergodic version) can be shown to ensure convergence of x^N with the desired convergence rate. In the following, we detail briefly our main results.
Let x_ϵ^* denote the solution of (13) and y_ϵ^* denote a solution of (15). Let X^* denote the set of minimizers of (2) and let us consider the following algorithm:

Algorithm 2 (Dual)

Choose a number of iterations N.
Set a point x^0 (as close as possible to X^*).
Set a starting point y^0 (as close as possible to y_ϵ^*).
Set $\epsilon=\epsilon(N)=\frac{|||A|||\cdot||x^0-x_\epsilon^*||_2}{N}$.
Set $\mathcal{A}=0$, $g=0$, $\bar{x}=0$ and $y=y^0$.
for $k=0$ to N **do**
  $a=\frac{1}{L}+\sqrt{\frac{1}{L^2}+\frac{2}{L}\mathcal{A}}$
  $v=\Pi_Y\left(y^0-g\right)$
  $z=\frac{\mathcal{A}y+av}{\mathcal{A}+a}$
  $y=\Pi_Y\left(z+\frac{\nabla\Psi_\epsilon(z)}{L}\right)$
  $\bar{x}=\bar{x}+ax(y)$ (cf. equation (17))
  $g=g-a\nabla\Psi_\epsilon(y)$
  $\mathcal{A}=\mathcal{A}+a$
end for
Set $\bar{x}^N=\frac{\bar{x}}{\mathcal{A}}$.

This algorithm can be shown to have the following properties.

Proposition 3 *$\bar{x}^N$ converges to the projection of x^0 onto the set of minimizers of (2).*

Proposition 4 *The worst case convergence rate is:*

$$\Psi(\bar{x}^N)-\Psi(x^*)\leq\frac{2|||A|||\cdot||x^0-x_\epsilon^*||_2\sqrt{D}}{N}. \quad (18)$$

Rate (18) is actually very similar to (12). It is thus natural to wonder if there is an interest in using this dual approach. Let us present some interesting aspects of this scheme:

- In the dual approach, the solution of the regularized problem is unique. This guarantees a certain stability of the iterates.
- We can show an additional convergence rate in norm to the regularized solution. Namely, for a fixed ϵ, we have for all k:

$$||\bar{x}^k-x_\epsilon^*||_2^2\leq\frac{D|||A|||}{\epsilon\cdot k^2} \quad (19)$$

- In practical experiments, model (13) with a small ϵ leads to slightly better SNR than model (2) for some restoration purposes in image processing.
- The practical convergence rates of the dual approach were better than those of the primal approach in all our experiments.

To conclude the theoretical part of this paper, let us precise that problem (3) can be solved with the same algorithms. However, it is preferable not to regularize the term $y\mapsto||y||_1$ which can be minimized using accelerated soft-thresholding algorithms [1, 10, 12].

5. Numerical results

In this section we present some comparisons for a problem of image zooming with impulse noise. To solve this problem, we simply set:

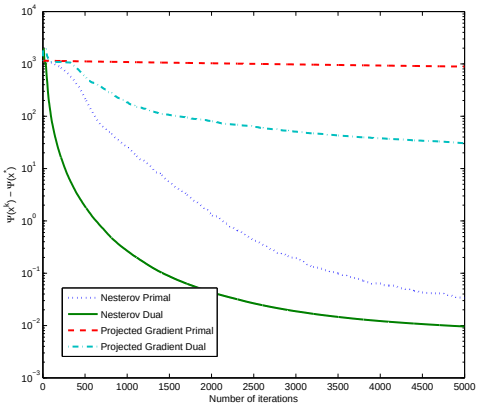


Figure 1: Cost function w.r.t. the number of iterations.

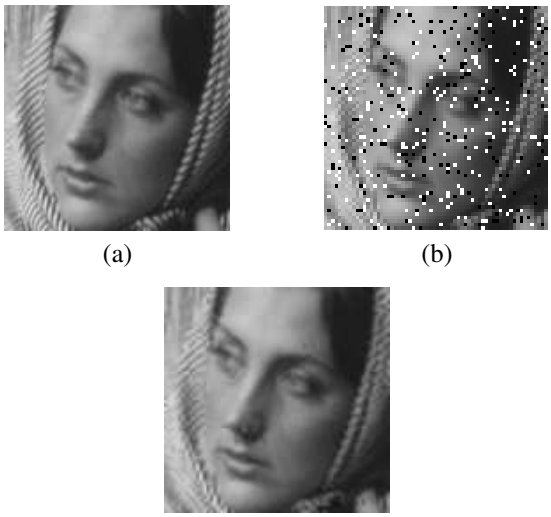
- D: convolution by a low-pass filter followed by a down sampling of factor d in the x and y directions.
- $p = 1$ (which is adapted to impulse noise).
- B: a redundant wavelet transform. We set B to be the Dual-Tree Complex Wavelet Tranform (DTCW) [11].

In that case $|||A|||^2$ can be computed explicitly. For the general case, let us point out that iterated power algorithms provide good approximations of $|||A^*A||| = |||A|||^2$.

In Figure 1, we chose $\epsilon = 0.045$ and $\mu = 10^{-5}$. This ensures that both methods lead to the same asymptotic accuracy (measured in terms of objective function). We can see that the dual approach seems to have a better behavior. For this problem reducing $\Psi(x^0) - \Psi(x^*)$ by a factor 10^3 is enough for visual purposes. The dual Nesterov approach requires 450 low cost iterations. The smoothing method proposed by Y. Nesterov requires 1700 iterations. The classical Cauchy steps requires much more than 5000 iterations to reach this goal. We can thus see the major improvement of Y. Nesterov's scheme on these problems. We carried out many other experiments which led to the same conclusion. Figure 2 shows the solution of the problem. The DTCW transform allows to retrieve thin details but slightly blurs the image. Further investigations will be led to address this issue.

6. Acknowledgments

The authors would like to thank the CS Compagny in Toulouse (France) for partial funding of this research work.



(a) (b)

(c)

Figure 2: Restoration of a down-sampled and noised image. (a) Original image, (b) down-sampled (by a factor 2) and noised image by 10% of "Salt & Pepper" noise and finally (c) result of the Nesterov dual approach.

References:

[1] A. Beck and M. Teboulle. Fast iterative shrinkagethresholding algorithm for linear inverse problems. *SIAM J. on Imaging Science*, to appear.

[2] A. Bermudez and C. Moreno. Duality methods for solving variational inequalities. *Comp. and Maths. with Appls.*, 7:43-58, 1981.

[3] R.J. Bruck. On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in hilbert space. *J. Math. Anal. Appl*, 61:159-164, 1977.

[4] J.F. Cai, R. Chan, Z.W. Shen, and L.X. Shen. Convergence analysis of tight framelet approach for missing data recoverys. *Advances in Computational Mathematics*, to appear.

[5] E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Inf. Theory*, 2006.

[6] A. Chambolle, R. Devore, N.Y. Lee, and B.J. Lucier. Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage. *IEEE Trans. Image Processing*, 7:319-335, 1998.

[7] G. Facciolo, A. Almansa, J.-F. Aujol, and Vicent Caselles. Irregular to regular sampling, denoising and deconvolution. *SIAM Journal on Multiscale Modeling and Simulation*, in press.

[8] A. Nemirovski. Information-based complexity of linear operator equations. *Journal of Complexity*, 8:153-175, 1992.

[9] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127-152, 2005.

[10] Y. Nesterov. Gradient methods for minimizing composite objective function. *CORE Discussion Paper 2007/76*, 2007.

[11] I. W. Selesnick, R. G. Baraniuk, and N. G. Kingsbury. The dual-tree complex wavelet transform. *IEEE Signal Processing Magazine*, 22(6), Nov. 2005.

[12] P. Weiss. *Algorithmes rapides d'optimisation convexe. Applications à la restauration d'images et à la détection de changements*. PhD thesis, Université de Nice Sophia Antipolis, Dec. 2008.

Image Inpainting Using a Fourth-Order Total Variation Flow

Carola-Bibiane Schönlieb* Andrea Bertozzi† Martin Burger‡ Lin He§

April 19, 2009

Abstract

We introduce a fourth-order total variation flow for image inpainting proposed in [5]. The well-posedness of this new inpainting model is discussed and its efficient numerical realization via an unconditionally stable solver developed in [15] is presented.

1 Introduction

An important task in image processing is the process of filling in missing parts of damaged images based on the information obtained from the surrounding areas. It is essentially a type of interpolation and is referred to as inpainting. Given an image f in a suitable Banach space of functions defined on $\Omega \subset \mathbb{R}^2$, an open and bounded domain, the problem is to reconstruct the original image u in the damaged domain $D \subset \Omega$, called inpainting domain. In the following we are especially interested in so called non-texture inpainting, i.e., the inpainting of structures, like edges and uniformly colored areas in the image, rather than texture.

In the pioneering works of Caselles et al. [6] (with the term disocclusion instead of inpainting) and Bertalmio et al. [2] partial differential equations have been first proposed for digital non-texture inpainting. In subsequent works variational models, originally derived for the tasks of image denoising, deblurring and segmentation, have been adopted to inpainting. The most famous variational inpainting model is the total variation (TV) model, cf. [8, 10, 13, 14]. Here, the inpainted image u is computed as a minimizer of the functional

$$\mathcal{J}(u) = |Du|(\Omega) + \frac{1}{2}\left\|\lambda(f-u)\right\|^2_{L^2(\Omega)},$$

where $|Du|(\Omega)$ is the total variation of u (cf. [1]), and λ is the indicator function of $\Omega \setminus D$ multiplied by a (large) constant, i.e., $\lambda(x) = \lambda_0 >> 1$ in $\Omega \setminus D$ and 0 in D. The corresponding steepest descent for the total variation inpainting model reads

$$u_t = -p + \lambda(f-u), \quad p \in \partial\,|Du|(\Omega), \tag{1}$$

where p is an element in the subdifferential of the total variation $\partial\,|Du|(\Omega)$. The steepest-descent approach is used to numerically compute a minimizer of $\mathcal{J}$, whereby it is iteratively solved until one is close enough to a minimizer of $\mathcal{J}$. For the numerical computation an element p of the subdifferential is approximated by the anisotropic diffusion $\nabla \cdot (\nabla u/|\nabla u|_\epsilon)$, where $|\nabla u|_\epsilon = \sqrt{|\nabla u|^2 + \epsilon}$.

Now, TV inpainting, while preserving edge information in the image, fails in propagating level lines (sets of image points with constant grayvalue) smoothly into the damaged domain, and in connecting edges over large gaps in particular. In an attempt to solve these issues from second order image diffusions, a number of third and fourth order diffusions have been suggested for image inpainting, e.g., [7, 9].

In this paper we present a fourth-order variant of total variation inpainting, called TV-H^{-1} inpainting. The inpainted image u of $f \in L^2(\Omega)$, shall evolve via

$$u_t = \Delta p + \lambda(f-u), \quad p \in \partial TV(u), \tag{2}$$

with

$$TV(u) = \begin{cases} |Du|(\Omega) & \text{if } |u(x)| \le 1 \text{ a.e. in } \Omega \\ +\infty & \text{otherwise.} \end{cases} \tag{3}$$

This inpainting approach has been proposed by Burger, He, and Schönlieb in [5] as a generalization of the sharp interface limit of Cahn-Hilliard inpainting [3, 4] to grayvalue images. The L^∞ bound in the definition (3) of the total variation functional $TV(u)$ is motivated by this sharp interface limit and is part of the technical setup, which made it easier to derive rigorous results for this scheme. A similar form of this higher-order TV approach already appeared in the context of decomposition and restoration of grayvalue images, see for example [12]. In the following we shall recall the main rigorous results obtained in [5], present an unconditionally stable solver for (2), and show a numerical example emphasizing the superiority of the fourth-order TV flow over the second-order one.

2 Well-Posedness of the Scheme

In contrast to its second-order analogue, the well-posedness of (2) strongly depends on the L^∞ bound introduced in (3).

*Department of Applied Mathematics and Theoretical Physics (DAMTP), Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WA, United Kingdom. Email: c.b.s.schonlieb@damtp.cam.ac.uk

†Department of Mathematics,UCLA (University of California Los Angeles), 405 Hilgard Avenue, Los Angeles, CA 90095-1555, USA. Email: bertozzi@math.ucla.edu

‡Institut für Numerische und Angewandte Mathematik, Fachbereich Mathematik und Informatik, Westfälische Wilhelms Universität (WWU) Münster, Einsteinstrasse 62, D 48149 Münster, Germany. Email: martin.burger@wwu.de

§Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenbergerstrasse 69, A-4040 Linz, Austria. Email: lin.he@oeaw.ac.at

This is because of the lack of maximum principles which, in the second-order case, guarantee the well-posedness for all smooth monotone regularizations of p.

The existence of a steady state for (2) is given by the following theorem.

Theorem 1 *[5, Theorem 1.4] Let $f \in L^2(\Omega)$. The stationary equation*

$$\Delta p + \lambda(f - u) = 0, \quad p \in \partial TV(u) \tag{4}$$

admits a solution $u \in BV(\Omega)$.

Results for the evolution equation (2) are a matter of future research. In particular it is highly desirable to achieve asymptotic properties of the evolution. Note that additionally to the fourth differential order, a difficulty in the convergence analysis of (2) is that it does not follow a variational principle.

3 Unconditionally Stable Solver

Motivated by the idea of convexity splitting schemes, e.g., [11], Bertozzi and Schönlieb propose in [15] the following time-stepping scheme for the numerical solution of (2):

$$\begin{array}{l} \frac{U_{k+1}-U_k}{\Delta t} + C_1\Delta\Delta U_{k+1} + C_2 U_{k+1} = C_1\Delta\Delta U_k \\ -\Delta(\nabla\cdot(\frac{\nabla U_k}{|\nabla U_k|_\epsilon})) + C_2 U_k + \lambda(f - U_k), \end{array} \tag{5}$$

with $C_1 > 1/\epsilon$, $C_2 > \lambda_0$. Here, U_k is the kth iterate of the time-discrete scheme, which approximates a solution u of the continuous equation at time $k\Delta t$, $\Delta t > 0$. It can be shown that (5) defines a numerical scheme that is unconditionally stable, and of order 2 in time, cf. [15].

4 Numerical Results

In Figure 1 a result of the TV-H^{-1} inpainting model computed via (5) and its comparison with the result obtained by the second order TV-L^2 inpainting model for a crop of the image is presented. The superiority of the fourth-order TV-H^{-1} inpainting model to the second order model with respect to the desired continuation of edges into the missing domain is clearly visible.

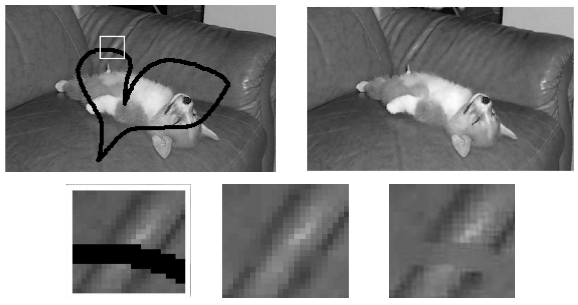


Figure 1: First row: TV-H^{-1} inpainting (2): u(1000) with $\lambda_0 = 10^3$. Second row: (l.) u(1000) with TV-H^{-1} inpainting, (r.) u(5000) with TV-L^2 inpainting (1)

Acknowledgments

CBS acknowledges the financial support provided by project WWTF Five senses-Call 2006, *Mathematical Methods for Image Analysis and Processing in the Visual Arts*, by the Wissenschaftskolleg (Graduiertenkolleg, Ph.D. program) of the Faculty for Mathematics at the University of Vienna (funded by FWF), and by the FFG project *Erarbeitung neuer Algorithmen zum Image Inpainting*, projectnumber 813610. Further, this publication is based on work supported by Award No. KUK-I1-007-43, made by King Abdullah University of Science and Technology (KAUST). For the hospitality and the financial support during parts of the preparation of this work, CBS thanks IPAM (UCLA), the US ONR Grant N000140810363, and the Department of Defense, NSF Grant ACI-0321917. The work of MB has been supported by the DFG through the project *Regularization with singular energies*.

References

[1] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Mathematical Monographs, Oxford University Press, 2000.

[2] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, *Image inpainting*, Computer Graphics, SIGGRAPH 2000, July, 2000.

[3] A. Bertozzi, S. Esedoglu, and A. Gillette, *Inpainting of Binary Images Using the Cahn-Hilliard Equation*. IEEE Trans. Image Proc. 16(1) pp. 285-291, 2007.

[4] A. Bertozzi, S. Esedoglu, and A. Gillette, *Analysis of a two-scale Cahn-Hilliard model for image inpainting*, Multiscale Modeling and Simulation, vol. 6, no. 3, pages 913-936, 2007.

[5] M. Burger, L. He, C. Schönlieb, *Cahn-Hilliard inpainting and a generalization for grayvalue images*, UCLA CAM report 08-41, June 2008.

[6] V. Caselles, J.-M. Morel, and C. Sbert, *An axiomatic approach to image interpolation*, IEEE Trans. Image Processing, 7(3):376386, 1998.

[7] T.F. Chan, S.H. Kang, and J. Shen, *Euler's elastica and curvature-based inpainting*, SIAM J. Appl. Math., Vol. 63, Nr.2, pp.564–592, 2002.

[8] T. F. Chan and J. Shen, *Mathematical models for local non-texture inpaintings*, SIAM J. Appl. Math., 62(3):10191043, 2001.

[9] T. F. Chan and J. Shen, *Non-texture inpainting by curvature driven diffusions (CDD)*, J. Visual Comm. Image Rep., 12(4):436449, 2001.

[10] T. F. Chan and J. Shen, *Variational restoration of non-flat image features: models and algorithms*, SIAM J. Appl. Math., 61(4):13381361, 2001.

[11] D. Eyre, *An Unconditionally Stable One-Step Scheme for Gradient Systems*, Jun. 1998, unpublished.

[12] S. Osher, A. Sole, and L. Vese. *Image decomposition and restoration using total variation minimization and the H -1 norm*, Multiscale Modeling and Simulation: A SIAM Interdisciplinary Journal, Vol. 1, Nr. 3, pp. 349-370, 2003.

[13] L. Rudin and S. Osher, *Total variation based image restoration with free local constraints*, Proc. 1st IEEE ICIP, 1:3135, 1994.

[14] L.I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D, Vol. 60, Nr.1-4, pp.259-268, 1992.

[15] C.-B. Schönlieb, and A. Bertozzi, *Unconditionally stable schemes for higher order inpainting*, in preparation.

Image Segmentation Through Efficient Boundary Sampling

Alex Chen[(1)], Todd Wittman[(1)] , Alexander Tartakovsky[(2)], Andrea Bertozzi[(1)]

(1) Department of Mathematics, University of California, Los Angeles, 520 Portola Plaza, Los Angeles, CA 90095
(2) Department of Mathematics, University of Southern California, 3620 S. Vermont Ave., KAP 108, Los Angeles, CA 90089
achen@math.ucla.edu, wittman@math.ucla.edu, tartakov@math.usc.edu, bertozzi@math.ucla.edu

Abstract:

This paper presents a combined geometric and statistical sampling algorithm for image segmentation inspired by a recently proposed algorithm for environmental sampling using autonomous robots [1].

1. Introduction

Segmentation is one of the most important problems in image processing. Partitioning an image into a small number of homogeneous regions highlights important features, allowing a user to analyze the image more easily. Applications include medical imaging, computer vision, and geospatial target detection. Image segmentation methods can be subdivided into region-based vs. edge-based methods. Region-based methods include the Mumford-Shah [2] and related Chan-Vese [3] methods which both involve energy minimization with a least squares fit of the data and a partition, between regions, whose length is minimized. Edge-based methods include the well-known image snakes [4] and Canny edge detector [5]. Other approaches to segmentation have also been effective. Statistical methods such as region competition rely on the fact that images have repetitive features that can be learned and exploited to obtain a segmentation [6]. A more recent fast statistical method called DistanceCut [7] is semi-supervised (the user identifies segments in each region) and is based on weighted distances and kernel density estimation.

All of these methods involve, at some level, sampling all the pixels in an image. For applications involving high-dimensional or large data sets, it makes sense to subsample the image. This is especially important for high resolution data where it can be prohibitive to perform calculations on every pixel in the image. The proposed segmentation method is designed for this kind of application and is based on ideas for cooperative environmental sampling with robotic vehicles.

The UUV-gas algorithm [8] utilizes robots that "walk" in a sinusoidal path along the boundary between two regions, changing directions as they cross from one region into another. This tracking method theoretically utilizes only those points that are near the boundary in question, resulting in substantial savings in run-time. The sinusoidal pattern has also been suggested as an efficient method for atomic force microscopy scanning [9]. Interestingly, the same idea of tracking is behind the sinusoidal walking pattern in ants following pheromone trails [10]. As with curve evolution methods in image processing, noise can cause problems, since the tracking is done as a local search. It was proposed [1] that the use of a change-point detection algorithm, e.g., Page's cumulative sum (CUSUM) algorithm [11] could improve tracking performance in noisy images. Testbed implementations of the boundary tracking algorithm exploiting change-point detection methods suggested that robots can indeed track boundaries efficiently in the presence of moderately intense noise [12]. We propose to adapt the above tracking algorithms to the problem of segmentation, with the goal of computational efficiency. Further improvements can be made that are not practical in the environmental tracking case. Many of these improvements are based on hypothesis testing for two regions, with the use of the CUSUM algorithm as a special case.

2. A two level sampling algorithm

The algorithm has two levels, namely a global searching method, which locates a boundary point, and a local sampling algorithm, which tracks the boundary using the global method as an initial point. Occasionally, if the tracker strays too far from the boundary, additional uses of the global algorithm are needed. We briefly discuss several options for the global search and then focus on the local sampling algorithm.

2.1 Global searching method – Locate an edge

Initialized at some point, the global search looks for some instance of the boundary. This can be done in a few ways. One method is simply to move out in a spiral pattern until a boundary point is detected (see Figure 1). However, if the boundary is small and far away from the initial point, it may be positioned between revolutions of the spiral and missed. Other options include deterministic paths that do not have the tendency to miss boundaries or stochastic paths using a random walk. These searching methods assume no prior knowledge of the boundary location, but they can be easily modified when some information is known. Another possibility is to implement a coarse segmentation of the data first and use the resulting boundary detection as an initialization for the local sampling. More

details on the last option are given later. Once a boundary point has been detected, the local sampling algorithm begins.

2.2 Local sampling algorithm – Track an edge

In the environmental tracking problem [1, 8], a robot tracks the boundary between two regions. The local sampling step is initialized at a point near the boundary, obtained from the global search. The robot then steers using a bang-bang steering controller, travelling in circular arcs, changing its direction of movement when it crosses into a different region.

It is relatively straightforward to adapt the algorithm for an image with domain Ω. As before, the problem is to find the boundary B between two regions, which will be labelled Ω_1 and Ω_2, so that $\Omega = \Omega_1 \cup \Omega_2 \cup B$ and $\Omega_1 \cap \Omega_2 = \emptyset$. Define an initial starting point $\vec{x}_0 = (x_0^1, x_0^2)$ for the boundary tracker and an initial value θ_0, representing the angle from the $+x^1$ direction, so that the initial direction vector is $(\cos\theta_0, \sin\theta_0)$. Also define the step size V and angular increment ω, which depend on estimates for image resolution and characteristics of the edge to be detected. In general, V is chosen smaller for greater detail, and ω is chosen smaller for straighter edges. A decision function between Ω_1 and Ω_2 must also be specified and has the following form:

$$d(\vec{x}) = \begin{cases} 1, & \text{if } \vec{x} \in \Omega_1, \\ 0, & \text{if } \vec{x} \in B, \\ -1, & \text{if } \vec{x} \in \Omega_2. \end{cases} \tag{1}$$

The simplest example is thresholding of the image intensity $I(\vec{x})$ at a given spatial location $\vec{x}$ (in the case of a grayscale image):

$$d(\vec{x}) = \begin{cases} 1, & \text{if } I(\vec{x}) > T, \\ 0, & \text{if } I(\vec{x}) = T, \\ -1, & \text{if } I(\vec{x}) < T, \end{cases} \tag{2}$$

where T is a fixed threshold value. Later in this section we use statistical information about prior points sampled along the path to modify $d(\vec{x})$. At each step k, the direction θ_k and current location $\vec{x}_k$ are updated recursively. Specifically, $\vec{x}_k = \vec{x}_{k-1} + V * (\cos\theta_{k-1}, \sin\theta_{k-1})$ and θ_k is updated according to the location of the new tracking head $\vec{x}_k$. A simple update for θ is the bang-bang steering controller, defined by

$$\theta_k = \theta_{k-1} + \omega d(\vec{x}_k). \tag{3}$$

An angle-correction modification [1] can be used for (3) if step k is a region crossing:

$$\theta_k = \theta_{k-1} + d(\vec{x}_k)(\bar{t}\omega - 2\theta_{ref})/2, \tag{4}$$

where $\bar{t}$ is the number of steps since the last region crossing, and θ_{ref} is a small fixed reference angle chosen based on the expected curvature of the edge being tracked.

One stopping condition for the tracking of finite regions is termination if the tracker comes within some range of the first boundary point detected, given some minimum number of iterations. Midpoints of line segments formed from region crossings are labelled boundary points.

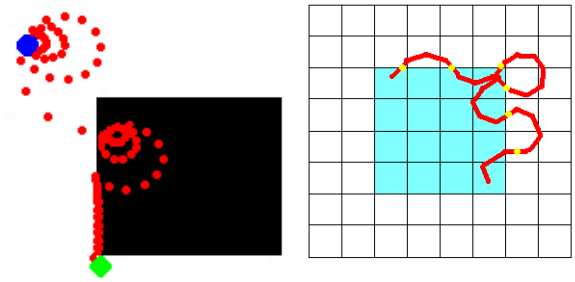


Figure 1: Left: Global search via a spiral-like pattern. The initial point is in blue, the final point (after a few iterations of local sampling) is in green, and the path is in red. Right: Basic procedure for the boundary tracking (local sampling) algorithm. The object is in cyan, the path of the tracking head is in red, and the detected boundary points are in yellow. Each small square represents one pixel. The tracker travels at fractional spatial values but samples at integral values.

While the local sampling method works well for clean images, it is susceptible to unavoidable errors in noisy images. Averaging readings from nearby pixels can minimize errors in the decision due to noise. In particular, sequential change-point detection methods are well-suited for detecting and tracking image edges in noise.

2.3 Decision algorithm

Change-point problems deal with detecting anomalies or changes in statistical behavior of data. The observations are obtained sequentially and, as long as their behavior is consistent with the normal state, one is content to let the process continue. If the state changes, then one is interested in detecting the change as soon as possible while minimizing false detections. More specifically, given a sequence of independent observations $s_1 = I(x_1), \ldots, s_n = I(x_n)$ and two probability density functions (pdf) f (pre-change) and g (post-change), determine whether there exists N such that the pdf of s_i is f for $i < N$ and g for $i \geq N$.

One of the most efficient change-point detection methods is the CUSUM algorithm proposed by Page in 1954 [11]. Write $Z_k = \log[g(s_k)/f(s_k)]$ for the log-likelihood ratio and define recursively

$$U_k = \max\left(U_{k-1} + Z_k, 0\right), \; U_0 = 0 \tag{5}$$

the CUSUM statistic and the corresponding stopping time $\tau = \min\{k \mid U_k \geq \overline{U}\}$, where $\overline{U}$ is a threshold controlling the false alarm rate. Then τ is a time of raising an alarm. In our applications, assuming that f is the pdf of the data in Ω_1 and g is the pdf in Ω_2, the value of τ may be interpreted as an estimate of the actual change-point, i.e., the boundary crossing from Ω_1 to Ω_2.

Note that if the pre-change and post-change densities f and g are completely specified, then the CUSUM algorithm performs optimally with respect to certain performance metrics [14]. However, in our applications these densities are usually unknown (while a Gaussian approximation may work well in certain scenarios). For this reason, the log-likelihood ratio Z_k in (5) should be replaced

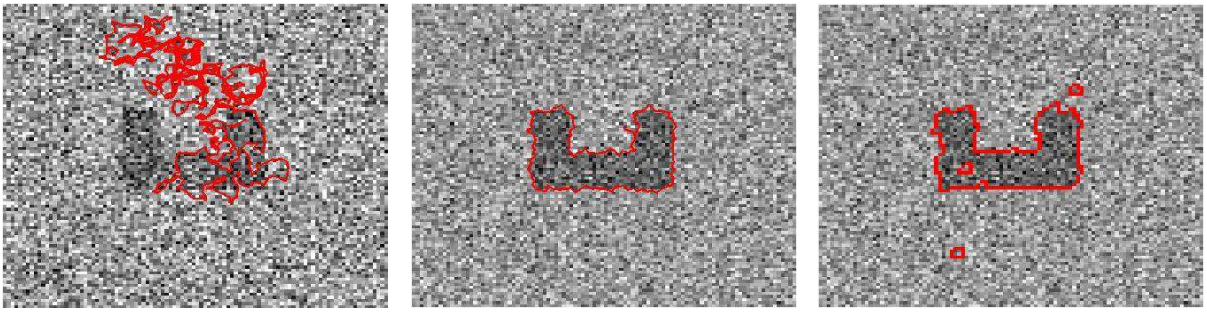


Figure 2: A 100 $\times$ 100 image was corrupted with additive Gaussian noise, N(0,0.5). Left: Boundary tracking without a change-point detection modification. Middle: Boundary tracking with the CUSUM algorithm. Right: Threshold dynamics [13].

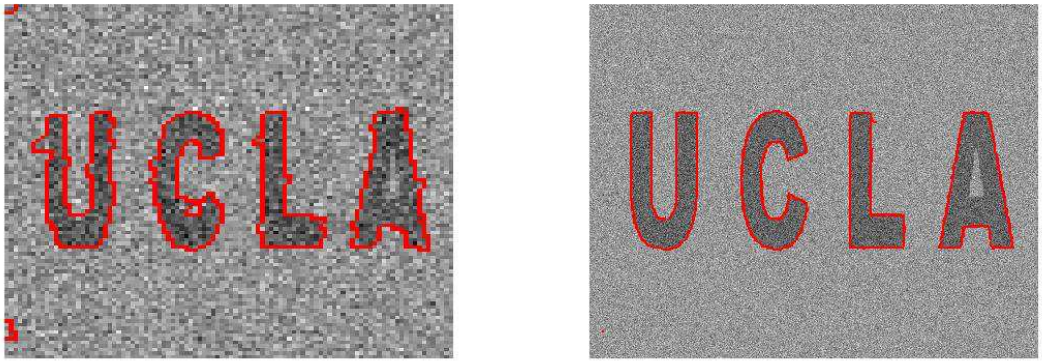


Figure 3: A hybrid level set – boundary tracking segmentation on a 1000 $\times$ 1000 image. Left: Initial segmentation by threshold dynamics. The image is subsampled by a factor of 10 on each axis. Right: Final segmentation by boundary tracking, using points from the connected components of the initial segmentation as starting points for trackers.

by a score function G_k sensitive to expected changes. Since we expect a change in the mean value, the appropriate score is $G_k = s_k - (\theta_1 + \theta_2)/2$, where θ_j is the mean of the previous observations s_i in Ω_j. The resulting score-based CUSUM test is not guaranteed to be optimal anymore. Note, however, that this score is optimal for Gaussian distributions (i.e., when sensor noise and residual clutter may be well approximated by Gaussian processes) and can be easily adjusted to cover any member of the exponential family of distributions (Bernoulli, Poisson, double exponential, etc.). For further details, see [15].

Changes from Ω_2 to Ω_1 can also be tracked in this manner. Analogously to (5) define recursively the decision statistic $L_k = \max(L_{k-1} - G_k, 0)$, $L_0 = 0$ and the stopping time $\tau = \min\{k \mid L_k \geq \overline{L}\}$, where G_k is the score introduced above, which is taken to be equal to Z_k if the distributions are known and where $\overline{L}$ is a threshold associated with a given false detection rate.

Only one of the statistics U_k or L_k is used at a time, i.e., when the tracking head is in Ω_1, the change-detection statistic U_k is used for detecting a transition to Ω_2. Similarly, when the tracking head is in Ω_2, only L_k is used for detecting a change to Ω_1. Once the tracking head enters a new region, the other statistic is used, initialized at 0.

Note that we have implicitly assumed that the intensity values on the path are independent observations. This assumption of independence is not entirely accurate, since the samples are taken from the tracking path, which is not a random sampling of an area. However, if noise levels are large, independence of observations is a relatively accurate assumption due to the spatial independence of noise, while if noise levels are small, the use of a change-detection algorithm is less important. Furthermore, the proposed score-based CUSUM tests are robust with respect to prior assumptions, including independence.

3. Boundary Tracking Examples

As mentioned above, one option for the global search is to run a coarse segmentation on a subsampled version of the image to obtain an initialization for the objects to be segmented. This "hybrid" method has an additional benefit of being able to detect mutiple objects and of giving a priori estimates for parameters in the decision function. The proposed two-stage hybrid boundary tracking algorithm that combines the UUV-gas algorithm with the CUSUM-based change-point detection identifies the true boundaries of an object accurately even in high levels of noise, as seen from Figure 2. The run-time and storage costs are minimal, compared to most other segmentation methods.

An example of a noisy image is shown in Figure 3. The original image is 1000 $\times$ 1000. Threshold Dynamics [13] was first applied to a heavily subsampled version (100 $\times$ 100) of the image. Then one pixel from each connected

component was taken as the starting point for a boundary tracker. An example using multispectral data is shown in Figure 4.

The hybrid method may be applied to more complicated images, but some problems arise. In the first step, when a coarse segmentation is applied to a subsampled image, small features may not be detected accurately. These small features will thus not be located by the boundary tracker either. Similarly, if some features are close in space, they may be placed in the same connected component class. In the boundary detection step, only one feature will thus be tracked. One solution is to use multiple initial points for each connected component. This will result in a decrease in efficiency but allow more objects to be tracked. Another problem is that different objects in the image may require different parameters to be chosen in the change-point detection algorithm. While some objects are detected accurately with certain parameters, often, some objects are not detected completely. Multichart CUSUM tests can be used effectively for this purpose.

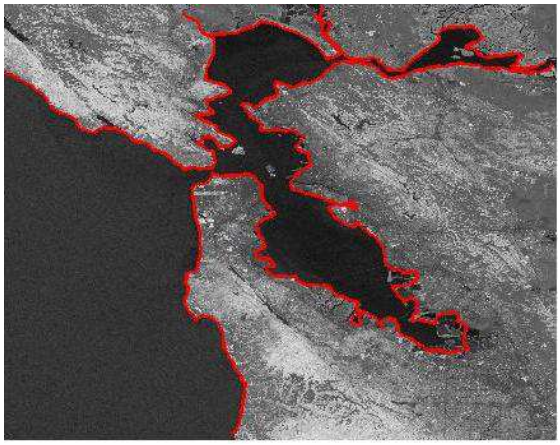


Figure 4: Boundary tracking of the San Francisco Bay coastline. A threshold of the Normalized Difference Vegetation Index (NDVI), commonly used for water detection [16], was taken as the decision function.

4. Discussion

The boundary tracking algorithm provides a fast alternative to many traditional segmentation methods due to its local nature. With the addition of a change-point detection method, the combined hybrid algorithm allows for accurate boundary tracking and, therefore, segmentation even in highly noisy images. Furthermore, the algorithm can operate efficiently even in data of large size or high resolution, scaling only with the size of the boundary rather than the size of the image. While presented as a novel segmentation method, the boundary tracking algorithm can also be used in conjunction with other segmentation methods in a two-stage algorithm.

Acknowledgments

The authors thank C. Bachmann, Z. Hu and V. Meija. This work was supported by the Department of Defense, ONR grant N000140810363, NSF ACI-0321917, ARO MURI 50363-MA-MUR.

References:

[1] Z. Jin and A. Bertozzi, "Environmental boundary tracking and estimation using multiple autonomous vehicles," *2007 46th IEEE Conference on Decision and Control*, pp. 4918–4923, December 2007.

[2] D. Mumford and J. Shah, "Optimal approximation by piecewise smooth functions and associated variational problems," *Communications on Pure and Applied Math*, vol. XLII, no. 5, pp. 577–684, July 1989.

[3] T. Chan and L. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 266–277, February 2001.

[4] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988.

[5] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, pp. 679–714, 1986.

[6] S. C. Zhu and A. L. Yuille, "Region competition: unifying snakes, region growing, and Bayes/MDL for multiband image segmentation," *IEEE Trans. on PAMI*, vol. 18, no. 9, pp. 884–900, 1996.

[7] X. Bai and G. Sapiro, "Distancecut: Interactive segmentation and matting of images and videos," *IEEE ICIP*, vol. 2, pp. 249–252, 2007.

[8] M. Kemp, A. L. Bertozzi, and D. Marthaler, "Multi-UUV perimeter surveillance," in *Proceedings of 2004 IEEE/OES Workshop on Autonomous Underwater Vehicles*, 2004, pp. 102–107.

[9] P. I. Chang and S. B. Andersson, "Smooth trajectories for imaging string-like samples in AFM: A preliminary study," in *2008 American Control Conference*, Seattle, Washington, June 2008.

[10] I. D. Couzin and N. R. Franks, "Self-organized lane formation and optimized traffic flow in army ants," *P. Roy. Soc. Lond. B. Bio.*, vol. 270, pp. 139–146, 2003.

[11] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1-2, pp. 100–115, June 1954.

[12] A. Joshi, T. Ashley, Y. Huang, and A. Bertozzi, "Experimental validation of cooperative environmental boundary tracking with on-board sensors." preprint.

[13] S. Esedoglu and Y. R. Tsai, "Threshold dynamics for the piecewise constant Mumford-Shah functional," *J. Comput. Phys.*, vol. 211, no. 1, pp. 367–384, 2006.

[14] G. Moustakides, "Optimal stopping times for detecting changes in distributions," *Annals of Statistics*, vol. 14, pp. 1379–1387, 1986.

[15] A. G. Tartakovsky, B. L. Rozovskii, R. Blažek, and H. Kim, "Detection of intrusions in information systems by sequential change-point methods," *Statistical Methodology*, vol. 3, no. 3, pp. 252–340, 2006.

[16] J. Rouse, R. Hass, J. Schell, and D. Deering, "Monitoring vegetation systems in the grain plains with ERTS," in *Third ERTS Symposium*, NASA SP-351 I, 1973, pp. 309–317.

SAMPTA'09

General Sessions

The Class of Bandlimited Functions with Unstable Reconstruction under Thresholding

Holger Boche and Ullrich J. Mönich

Technische Universität Berlin, Heinrich-Hertz-Chair for Mobile Communications,
Einsteinufer 25, 10578 Berlin, Germany.
{holger.boche,ullrich.moenich}@mk.tu-berlin.de

Abstract:
The reconstruction of $\mathcal{PW}_\pi^1$-functions by sampling series is not possible in general if the samples are disturbed by the non-linear threshold operator which sets all samples whose absolute value is smaller than some threshold to zero. In this paper we characterize the set of functions for which the sampling series diverges as the threshold goes to zero and show that this set is a residual set.

1. Notation

Before we start our discussion, we introduce some notations and definitions [4]. Let $\hat{f}$ denote the Fourier transform of a function f, where $\hat{f}$ is to be understood in the distributional sense. $L^p(\mathbb{R})$, $1 \leq p < \infty$, is the space of all measurable, pth-power Lebesgue integrable functions on $\mathbb{R}$, with the usual norm $\|\cdot\|_p$, and $L^\infty(\mathbb{R})$ is the space of all measurable functions for which the essential supremum norm $\|\cdot\|_\infty$ is finite.
For $\sigma > 0$ and $1 \leq p \leq \infty$ we denote by $\mathcal{PW}_\sigma^p$ the Paley-Wiener space of functions f with a representation $f(z) = 1/(2\pi) \int_{-\sigma}^{\sigma} g(\omega)\, \mathrm{e}^{\mathrm{i}z\omega}\ \mathrm{d}\omega$, $z \in \mathbb{C}$, for some $g \in L^p(-\sigma,\sigma)$. If $f \in \mathcal{PW}_\sigma^p$ then $g(\omega) = \hat{f}(\omega)$. The norm for $\mathcal{PW}_\sigma^p$, $1 \leq p < \infty$, is given by $\|f\|_{\mathcal{PW}_\sigma^p} = (1/(2\pi) \int_{-\sigma}^{\sigma} |\hat{f}(\omega)|^p\ \mathrm{d}\omega)^{1/p}$.
Furthermore, we need the threshold operator. For complex numbers $z \in \mathbb{C}$, the threshold operator κ_δ, $\delta > 0$, is defined by

$$\kappa_\delta z = \begin{cases} z & |z| \geq \delta \\ 0 & |z| < \delta. \end{cases}$$

For continuous functions $f : \mathbb{R} \to \mathbb{C}$, we define the threshold operator Θ_δ, $\delta > 0$, pointwise, i.e., $(\Theta_\delta f)(t) = \kappa_\delta f(t)$, $t \in \mathbb{R}$.

2. Motivation

A well known fact [1, 2, 3] about the convergence behavior of the Shannon sampling series for $f \in \mathcal{PW}_\pi^1$ is expressed by the following theorem.

Theorem (Brown). *For all $f \in \mathcal{PW}_\pi^1$ and $T > 0$ fixed we have*

$$\lim_{N\to\infty} \max_{t\in[-T,T]} \left| f(t) - \sum_{k=-N}^{N} f(k) \frac{\sin(\pi(t-k))}{\pi(t-k)} \right| = 0. \tag{1}$$

This theorem plays a fundamental role in applications, because it establishes the uniform convergence on compact subsets of $\mathbb{R}$ for a large class of functions, namely $\mathcal{PW}_\pi^1$, which is the largest space within the scale of Paley-Wiener spaces.
The truncation of the series in (1) is done in the domain of the function f because only the samples $f(k)$, $k = -N, \ldots, N$ are taken into account. In contrast, it is also possible to control the truncation of the series in the codomain of f by considering only the samples $f(k)$, $k \in \mathbb{Z}$, whose absolute value is larger than or equal to some threshold $\delta > 0$. This leads to the approximation formula

$$(A_\delta f)(t) = \sum_{\substack{k=-\infty \\ |f(k)|\geq\delta}}^{\infty} f(k) \frac{\sin(\pi(t-k))}{\pi(t-k)}. \tag{2}$$

Since $f \in \mathcal{PW}_\pi^1$ we have $\lim_{t\to\infty} f(t) = 0$ by the Riemann-Lebesgue lemma, and it follows that the series in (2) has only finitely many summands, which implies $A_\delta f \in \mathcal{PW}_\pi^2 \subset \mathcal{PW}_\pi^1$. In general, $A_\delta f$ is only an approximation of f, and we want the function $A_\delta f$ to be close to f if δ is sufficiently small.
The operator A_δ has several properties which complicate its analysis. A_δ, $\delta > 0$, is non-linear. Furthermore, for each $\delta > 0$, the operator $A_\delta : (\mathcal{PW}_\pi^1, \|\cdot\|_{\mathcal{PW}_\pi^1}) \to (\mathcal{PW}_\pi^1, \|\cdot\|_\infty)$ is discontinuous. This implies that $A_\delta : (\mathcal{PW}_\pi^1, \|\cdot\|_{\mathcal{PW}_\pi^1}) \to (\mathcal{PW}_\pi^1, \|\cdot\|_{\mathcal{PW}_\pi^1})$ is discontinuous. For some $f \in \mathcal{PW}_\pi^1$, the operator A_δ is also discontinuous with respect to δ.
Of course (2) can be written as

$$\sum_{k=-\infty}^{\infty} (\Theta_\delta f)(k) \frac{\sin(\pi(t-k))}{\pi(t-k)}, \tag{3}$$

where Θ_δ denotes the threshold operator. Wireless sensor networks are one possible application where the threshold operator Θ_δ and the series (3) are important. The sensors sample some bandlimited signal in space and time and then transmit the samples to the receiver. In order to save energy, it is common to let the sensors transmit only if the absolute value of the current sample exceeds some threshold $\delta > 0$. Thus, the receiver has to reconstruct the signal by using only the samples whose absolute value is larger than or equal to the threshold δ.

In addition to the sensor network scenario, the threshold operator can be used to model non-linearities in many other applications. For example, due to its close relation to the quantization operator, the threshold operator can be employed to analyze the effects of quantization in analog to digital conversion.

3. Problem Formulation and Main Result

Since the series in (2) uses all "important" samples of the function, i.e., all samples that are larger than or equal to δ, one could expect $A_\delta f$ to have an approximation behavior similar to the Shannon sampling series. In particular the approximation error should decrease as the threshold δ goes to zero. But, we will see that $A_\delta f$ exhibits a significantly different behavior.
In this paper we are interested in the structure of the set

$$\mathcal{D}_1 = \{f \in \mathcal{PW}_\pi^1 : \limsup_{\delta\to 0} |(A_\delta f)(t)| = \infty \,\forall\, t \in \mathbb{R}\setminus\mathbb{Z}\},$$

i.e., in the structure of the set of functions for which the approximation error $|f(t) - (A_\delta f)(t)|$ grows arbitrarily large for all $t \in \mathbb{R}\setminus\mathbb{Z}$ as $\delta \to 0$.

Remark 1. The analysis of the operator A_δ is difficult because it is non-linear and discontinuous, and therefore the standard theorems of functional analysis, like the Banach-Steinhaus theorem, cannot be used.

For the further discussion we need the following concepts from metric spaces [5]. A subset M of a metric space X is said to be nowhere dense in X if the closure $[M]$ does not contain a non-empty open set of X. M is said to be of the first category (or meager) if M is the countable union of sets each of which is nowhere dense in X. M is said to be of the second category (or nonmeager) if is not of the first category. The complement of a set of the first category is called a residual set. Sets of first category may be considered as "small". According to Baire's theorem [5] we have that in a complete metric space, the residual set is dense and a set of the second category. One property that shows the richness of residual sets is the following: The countable intersection of residual sets is always a residual set. In particular we will use the following fact in our proof. In a complete metric space an open and dense set is a residual set because its complement is nowhere dense.
Theorem 1 will show that the set $\mathcal{D}_1$ it is a residual set. Thus the threshold operator destroys the good reconstruction behavior of the Shannon sampling series for "almost all" functions in $\mathcal{PW}_\pi^1$.

4. Proof of the Main Result

In addition to the threshold operator that sets all samples whose absolute value is smaller than δ to zero, we consider the threshold operator that sets all samples whose absolute value is smaller than or equal to δ to zero. This operator gives rise to the sampling series

$$(\bar{A}_\delta f)(t) := \sum_{\substack{k=-\infty \\ |f(k)|>\delta}}^{\infty} f(k)\frac{\sin(\pi(t-k))}{\pi(t-k)} \qquad (4)$$

and the set

$$\mathcal{D}_2 = \{f \in \mathcal{PW}_\pi^1 : \limsup_{\delta\to 0} |(\bar{A}_\delta f)(t)| = \infty \,\forall\, t \in \mathbb{R}\setminus\mathbb{Z}\}.$$

Both threshold operators and thus A_δ and $\bar{A}_\delta$ are meaningful in practical applications, and one would expect the difference being not important. However, as we will see, $\bar{A}_\delta$ can be analyzed more easily.
For $\hat{t} \in \mathbb{R}\setminus\mathbb{Z}$ we furthermore define the sets

$$\mathcal{D}_1(\hat{t}) = \{f \in \mathcal{PW}_\pi^1 : \limsup_{\delta\to 0} |(A_\delta f)(\hat{t})| = \infty\}$$

and

$$\mathcal{D}_2(\hat{t}) = \{f \in \mathcal{PW}_\pi^1 : \limsup_{\delta\to 0} |(\bar{A}_\delta f)(\hat{t})| = \infty\}.$$

Lemma 1 shows that we do not have to distinguish between the sets D_1 and $D_1(\hat{t})$ and between D_2 and $D_2(\hat{t})$.

Lemma 1. *For all $\hat{t} \in \mathbb{R}\setminus\mathbb{Z}$ we have $D_1 = D_1(\hat{t})$ and $D_2 = D_2(\hat{t})$.*

Proof. The inclusion $D_1 \subset D_1(\hat{t})$ is obvious. It remains to show that $D_1(\hat{t}) \subset D_1$. Let $f \in D_1(\hat{t})$. For all $t_1 \in \mathbb{R}\setminus\mathbb{Z}$ and $\delta > 0$ a short calculation shows that

$$\left|\frac{1}{\sin(\pi t_1)}(A_\delta f)(t_1) - \frac{1}{\sin(\pi\hat{t})}(A_\delta f)(\hat{t})\right|$$
$$\le \|f\|_{\mathcal{PW}_\pi^1}\frac{|\hat{t}-t_1|}{\pi}\sum_{k=-\infty}^{\infty}\frac{1}{|t_1-k||\hat{t}-k|} = C_1(t_1,\hat{t},f),$$

where $C_1(t_1,\hat{t},f) < \infty$ is a constant that only depends on t_1, $\hat{t}$, and f. It follows that

$$|(A_\delta f)(t_1)| \ge |(A_\delta f)(\hat{t})|\left|\frac{\sin(\pi t_1)}{\sin(\pi\hat{t})}\right| - C_2(t_1,\hat{t},f). \quad (5)$$

Taking the limit superior on both sides of (5) gives

$$\limsup_{\delta\to 0} |(A_\delta f)(t_1)| = \infty. \qquad (6)$$

Since (6) is valid for all $t_1 \in \mathbb{R}\setminus\mathbb{Z}$, it follows that $f \in \mathcal{D}_1$. The same calculation shows that $D_2 = D_2(\hat{t})$. □

According to Lemma 1 it is sufficient to restrict the analysis to the sets $\mathcal{D}_1(\hat{t})$ and $\mathcal{D}_2(\hat{t})$ for some $\hat{t} \in \mathbb{R}\setminus\mathbb{Z}$. Furthermore, we can concentrate on one of both sets, because of the following lemma.

Lemma 2. *We have $\mathcal{D}_1 = \mathcal{D}_2$.*

Proof. Let $f \in \mathcal{D}_2(\hat{t})$ be arbitrary but fixed. By the definition of $\mathcal{D}_2(\hat{t})$, we have $\limsup_{\delta\to 0}|(\bar{A}_\delta f)(\hat{t})| = \infty$. Thus, for every $M > 0$ there exists a $\delta_M > 0$ such that $|(\bar{A}_{\delta_M} f)(\hat{t})| > M$. Let $\mathcal{T}(M) = \{k \in \mathbb{Z} : |f(k)| > \delta_M\}$ and $\underline{f}_M = \min_{k\in\mathcal{T}(M)}|f(k)|$. Then it follows that $\underline{f}_M > \delta_M$. Moreover, for all δ with $\underline{f}_M > \delta > \delta_M$ we have

$$(A_\delta f)(\hat{t}) = \sum_{\substack{k=-\infty \\ |f(k)|\ge\delta}}^{\infty} f(k)\frac{\sin(\pi(\hat{t}-k))}{\pi(\hat{t}-k)}$$
$$= \sum_{\substack{k=-\infty \\ |f(k)|>\delta_M}}^{\infty} f(k)\frac{\sin(\pi(\hat{t}-k))}{\pi(\hat{t}-k)} = (\bar{A}_{\delta_M} f)(\hat{t}).$$

Consequently,

$$\sup_{\delta>0}|(A_\delta f)(\hat{t})| > M. \tag{7}$$

Since (7) is valid for all $M > 0$, it follows that $\sup_{\delta>0}|(A_\delta f)(\hat{t})| = \infty$, and, as a consequence, $\limsup_{\delta\to 0}|(A_\delta f)(\hat{t})| = \infty$, because $|(A_\delta f)(\hat{t})| < \infty$ for all $\delta > 0$. This shows that $f \in \mathcal{D}_1(\hat{t})$, which implies that $\mathcal{D}_2(\hat{t}) \subset \mathcal{D}_1(\hat{t})$. The converse $\mathcal{D}_2(\hat{t}) \supset \mathcal{D}_1(\hat{t})$ is shown similarly. Hence $\mathcal{D}_1(\hat{t}) = \mathcal{D}_2(\hat{t})$, and the statement $\mathcal{D}_1 = \mathcal{D}_2$ follows from Lemma 1. □

In order to prove our main result, we need the important Lemma 3.

Lemma 3. *For all $M \in \mathbb{N}$ and $\hat{t} \in \mathbb{R} \setminus \mathbb{Z}$,*

$$\mathcal{D}_2(\hat{t}, M) = \{f \in \mathcal{PW}_\pi^1 : \sup_{\delta>0}|(\bar{A}_\delta f)(\hat{t})| > M\}$$

is a residual set.

Proof. Let $M \in \mathbb{N}$ and $\hat{t} \in \mathbb{R} \setminus \mathbb{Z}$ be arbitrary but fixed. First, we show that $\mathcal{D}_2(\hat{t}, M)$ is an open set. Let $f_1 \in \mathcal{D}_2(\hat{t}, M)$ be arbitrary. We have to show that there exists an $\epsilon > 0$ such that, given any $f \in \mathcal{PW}_\pi^1$ with $\|f - f_1\|_{\mathcal{PW}_\pi^1} < \epsilon$, $f \in \mathcal{D}_2(\hat{t}, M)$. By assumption, there exists a $\delta_M > 0$ such that

$$|(\bar{A}_{\delta_M} f_1)(\hat{t})| > M.$$

Furthermore, let $\mathcal{T}(M) = \{k \in \mathbb{Z} : |f_1(k)| > \delta_M\}$ and $\underline{f}_{1,M} = \min_{k\in\mathcal{T}(M)}|f_1(k)|$. Next, we choose $\tilde{\delta}_M = \delta_M + (\underline{f}_{1,M} - \delta_M)/2$. Then we have that

$$\{k \in \mathbb{Z} : |f_1(k)| > \tilde{\delta}_M\} = \mathcal{T}(M). \tag{8}$$

We choose

$$\tilde{\epsilon} < \min\left(\frac{|(\bar{A}_{\tilde{\delta}_M} f_1)(\hat{t})| - M}{|\mathcal{T}(M)|}, \tilde{\delta}_M - \delta_M\right). \tag{9}$$

For all $f \in \mathcal{PW}_\pi^1$ with $\|f_1 - f\|_{\mathcal{PW}_\pi^1} < \tilde{\epsilon}$ we have $|f_1(k) - f(k)| < \tilde{\epsilon}$, $k \in \mathbb{Z}$. It follows, for all $k \in \mathbb{Z}$ with $|f(k)| > \tilde{\delta}_M$, that

$$|f_1(k)| \geq |f(k)| - |f(k) - f_1(k)| > \tilde{\delta}_M - \tilde{\epsilon} > \delta_M,$$

i.e., $k \in \mathcal{T}(M)$. Conversely, $k \in \mathcal{T}(M)$ implies $f_1(k) \geq \underline{f}_{1,M}$, and it follows that

$$\begin{aligned}|f(k)| &\geq |f_1(k)| - |f(k) - f_1(k)| > \underline{f}_{1,M} - \tilde{\epsilon} \\ &> \underline{f}_{1,M} - \tilde{\delta}_M + \delta_M = \tilde{\delta}_M.\end{aligned}$$

Thus we have

$$\{k \in \mathbb{Z} : |f(k)| > \tilde{\delta}_M\} = \mathcal{T}(M). \tag{10}$$

Moreover, using (8) and (10), we obtain that

$$\begin{aligned}&|(\bar{A}_{\tilde{\delta}_M} f)(\hat{t}) - (\bar{A}_{\tilde{\delta}_M} f_1)(\hat{t})| \\ &= \left|\sum_{\substack{k=-\infty \\ |f(k)|>\tilde{\delta}_M}}^{\infty} f(k)\frac{\sin(\pi(\hat{t}-k))}{\pi(\hat{t}-k)} - \sum_{\substack{k=-\infty \\ |f_1(k)|>\delta_M}}^{\infty} f_1(k)\frac{\sin(\pi(\hat{t}-k))}{\pi(\hat{t}-k)}\right| \\ &\leq \sum_{k\in\mathcal{T}(M)} |f_1(k) - f(k)|\left|\frac{\sin(\pi(\hat{t}-k))}{\pi(\hat{t}-k)}\right| \leq \tilde{\epsilon}|\mathcal{T}(M)|\end{aligned}$$

and consequently

$$|(\bar{A}_{\tilde{\delta}_M} f)(\hat{t})| \geq |(\bar{A}_{\tilde{\delta}_M} f_1)(\hat{t})| - \tilde{\epsilon}|\mathcal{T}(M)| > M,$$

where the last inequality is due to (9). Therefore

$$\sup_{\delta>0}|(\bar{A}_\delta f)(\hat{t})| > M,$$

i.e., $f \in \mathcal{D}_2(\hat{t}, M)$, for all $f \in \mathcal{PW}_\pi^1$ with $\|f_1 - f\|_{\mathcal{PW}_\pi^1} < \tilde{\epsilon}$.

Second, we show that $\mathcal{D}_2(\hat{t}, M)$ is dense in $\mathcal{PW}_\pi^1$. Let $f \in \mathcal{PW}_\pi^1$ be arbitrary. We have to show that for every $\epsilon > 0$ there exists a $f_\epsilon \in \mathcal{D}_2(\hat{t}, M)$ such that $\|f - f_\epsilon\|_{\mathcal{PW}_\pi^1} < \epsilon$. Let $\epsilon > 0$ be arbitrary but fixed. Since $\mathcal{PW}_\pi^2$ is dense in $\mathcal{PW}_\pi^1$, there exists a $f_\epsilon^{(1)} \in \mathcal{PW}_\pi^2$ with

$$\|f - f_\epsilon^{(1)}\|_{\mathcal{PW}_\pi^1} < \frac{\epsilon}{3}. \tag{11}$$

Moreover, there exists a $f_\epsilon^{(2)} \in \mathcal{PW}_\pi^2$ such that $f_\epsilon^{(2)}(k) \neq 0$ only for finitely many $k \in \mathbb{Z}$ and

$$\|f_\epsilon^{(1)} - f_\epsilon^{(2)}\|_{\mathcal{PW}_\pi^1} < \frac{\epsilon}{3}. \tag{12}$$

Let N denote the smallest natural number such that $N > \hat{t}$ and $f_\epsilon^{(2)}(k) = 0$ for all $|k| > N$. Furthermore, let $\mathcal{T}_2 = \{k \in \mathbb{Z} : |f_\epsilon^{(2)}(k)| \neq 0\}$ and $\underline{f}_\epsilon^{(2)} = \min_{k\in\mathcal{T}_2}|f_\epsilon^{(2)}(k)|$. For $0 < \eta < 1$ and $L \in \mathbb{N}$, $L > N$, consider the functions h and g defined by

$$h(t, \eta, L) := \sum_{k=-2L+1}^{2L-1} h(k, \eta, L)\frac{\sin(\pi(t-k))}{\pi(t-k)},$$

where

$$h(k,\eta,L) = \begin{cases}(-1)^k(2(1-\eta)+\frac{1-\eta}{L}k), & -2L<k<-L,\\ (-1)^k(1-\eta), & -L\leq k<0,\\ (-1)^k, & 0\leq k\leq L,\\ (-1)^k(2-\frac{1}{L}k), & L<k<2L,\end{cases}$$

and

$$\begin{aligned}g(t,\eta,L) :=& h(t,\eta,L) - \underbrace{\sum_{k=0}^{N}\frac{(-1)^k\sin(\pi(t-k))}{\pi(t-k)}}_{=:u_1} \\ &- \underbrace{\sum_{k=-N}^{-1}(1-\eta)\frac{(-1)^k\sin(\pi(t-k))}{\pi(t-k)}}_{=:u_2}.\end{aligned}$$

Note that $g(k, \eta, L) = 0$ for $|k| \leq N$. We have

$$\begin{aligned}&\|g(t,\eta,L)\|_{\mathcal{PW}_\pi^1} \\ &\quad\leq \|h(\,\cdot\,,\eta,L)\|_{\mathcal{PW}_\pi^1} + \|u_1\|_{\mathcal{PW}_\pi^1} + \|u_2\|_{\mathcal{PW}_\pi^1}.\end{aligned} \tag{13}$$

The norm $\|u_1\|_{\mathcal{PW}_\pi^1}$ is upper bounded by

$$\|u_1\|_{\mathcal{PW}_\pi^1} < \frac{\pi}{2} + \log(N+1), \tag{14}$$

because

$$\|u_1\|_{\mathcal{PW}^1_\pi} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\sum_{k=0}^{N} \mathrm{e}^{-i\omega k}(-1)^k\right| \,\mathrm{d}\omega$$
$$= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{1-\mathrm{e}^{i\omega(N+1)}}{1-\mathrm{e}^{i\omega}}\right|\mathrm{d}\omega = \frac{1}{\pi}\int_0^{\pi}\left|\frac{\sin(\frac{N+1}{2}\omega)}{\sin(\frac{\omega}{2})}\right|\mathrm{d}\omega$$
$$\leq \int_0^{\pi}\frac{\left|\sin(\frac{N+1}{2}\omega)\right|}{\omega}\,\mathrm{d}\omega = \int_0^{N+1}\frac{\left|\sin(\frac{\pi}{2}\omega)\right|}{\omega}\,\mathrm{d}\omega$$
$$\leq \int_0^{1}\frac{\sin(\frac{\pi}{2}\omega)}{\omega}\,\mathrm{d}\omega + \int_1^{N+1}\frac{1}{\omega}\,\mathrm{d}\omega < \frac{\pi}{2} + \log(N+1).$$

A similar calculation gives

$$\|u_2\|_{\mathcal{PW}^1_\pi} < \frac{\pi}{2} + \log(N). \tag{15}$$

In addition we have $\|h(\,\cdot\,,0,L)\|_{\mathcal{PW}^1_\pi} \leq 3$, which can be proven easily, and $\lim_{\eta\to 0}\|h(\,\cdot\,,\eta,L) - h(\,\cdot\,,0,L)\|_{\mathcal{PW}^1_\pi} = 0$. Therefore, there exists an $0 < \eta_0(L) < 1$ such that

$$\|h(\,\cdot\,,\eta_0(L),L)\|_{\mathcal{PW}^1_\pi} < 4. \tag{16}$$

Combining (13)–(16) gives, that for all $L \in \mathbb{N}$, $L > N$ there exists an $0 < \eta_0(L) < 1$ such that

$$\|g(\,\cdot\,,\eta_0(L),L)\|_{\mathcal{PW}^1_\pi} < 4 + \pi + 3\log(N+1) =: C_3.$$

It is important that the constant C_3 does not depend on L. Next, we analyze

$$G_\epsilon(t,L) = f_\epsilon^{(2)}(t) + \mu g(t,\eta_0(L),L),$$

where $\mu > 0$ is some real number that satisfies $\mu < \min(\epsilon/(3C_3), \underline{f}_\epsilon^{(2)})$. By the choice of μ we have

$$\|f_\epsilon^{(2)} - G_\epsilon(\,\cdot\,,L)\|_{\mathcal{PW}^1_\pi} = \mu C_3 < \frac{\epsilon}{3} \tag{17}$$

for all $L > N$. Combining (11), (12), and (17), we see that

$$\|f - G_\epsilon(\,\cdot\,,L)\|_{\mathcal{PW}^1_\pi} < \epsilon \tag{18}$$

for all $L > N$, i.e., $G_\epsilon(\,\cdot\,,L)$ lies in the ϵ-ball around f. Furthermore, for any $L > N$ we can find a $\delta_0(L)$ that fulfills

$$\max\left((1-\eta_0(L))\mu, \left(1-\frac{1}{L}\right)\mu\right) < \delta_0(L) < \mu.$$

Since $\delta_0(L) < \underline{f}_\epsilon^{(2)}$, by the definition of μ, it follows that

$$(\bar{A}_{\delta_0(L)}G_\epsilon(\,\cdot\,,L))(\hat{t})$$
$$= \sum_{\substack{k=-N\\ |G_\epsilon(k,L)|>\delta_0(L)}}^{N} G_\epsilon(k,L)\frac{\sin(\pi(\hat{t}-k))}{\pi(\hat{t}-k)}$$
$$+ \sum_{\substack{k=N+1\\ |G_\epsilon(k)|>\delta_0(L)}}^{L} G_\epsilon(k,L)\frac{\sin(\pi(\hat{t}-k))}{\pi(\hat{t}-k)}$$
$$= \sum_{k=-N}^{N} f_\epsilon^{(2)}(k)\frac{\sin(\pi(\hat{t}-k))}{\pi(\hat{t}-k)} + \mu\sum_{k=N+1}^{L}\frac{(-1)^k\sin(\pi(\hat{t}-k))}{\pi(\hat{t}-k)}$$
$$= f_\epsilon^{(2)}(\hat{t}) + \mu\frac{\sin(\pi\hat{t})}{\pi}\sum_{k=N}^{L}\frac{1}{\hat{t}-k}.$$

Observing that $N - \hat{t} > 0$, we obtain

$$\left|\sum_{k=N}^{L}\frac{1}{\hat{t}-k}\right| = \sum_{k=0}^{L-N}\frac{1}{k+N-\hat{t}}$$
$$\geq \sum_{k=0}^{L-N}\int_k^{k+1}\frac{1}{\tau+N-\hat{t}}\,\mathrm{d}\tau$$
$$= \int_0^{L-N+1}\frac{1}{\tau+N-\hat{t}}\,\mathrm{d}\tau$$
$$> \log\left(\frac{L-\hat{t}}{N-\hat{t}}\right),$$

and consequently

$$|(\bar{A}_{\delta_0(L)}G_\epsilon(\,\cdot\,,L))(\hat{t})|$$
$$\geq \mu\frac{|\sin(\pi\hat{t})|}{\pi}\log\left(\frac{L-\hat{t}}{N-\hat{t}}\right) - |f_\epsilon^{(2)}(\hat{t})|. \tag{19}$$

The right-hand side of (19) can be made arbitrarily large by choosing L large. Let $L_1 > N$ be the smallest L such that the right hand side of (19) is larger than M. It follows that $f_\epsilon(t) = G_\epsilon(t,L_1)$ is the desired function, because $\sup_{\delta>1}|(\bar{A}_\delta f_\epsilon)(\hat{t})| \geq |(\bar{A}_{\delta_0(L_1)}f_\epsilon)(\hat{t})| > M$, i.e., $f_\epsilon \in \mathcal{D}_2(\hat{t},M)$, and because $\|f - f_\epsilon\|_{\mathcal{PW}^1_\pi} < \epsilon$, according to (18). □

Theorem 1. *$\mathcal{D}_1$ and $\mathcal{D}_2$ are residual sets.*

Proof. Since $\mathcal{D}_2 = \mathcal{D}_1$, by Lemma 2, it is sufficient to show that $\mathcal{D}_2$ is a residual set.
Let $\hat{t} \in \mathbb{R}\setminus\mathbb{Z}$ be arbitrary but fixed. We have

$$\mathcal{D}_2(\hat{t}) = \bigcap_{M\in\mathbb{N}} \mathcal{D}_2(\hat{t},M).$$

From Lemma 3 we know that all $\mathcal{D}_2(\hat{t},M)$, $M \in \mathbb{N}$, are residual sets. It follows that $\mathcal{D}_2(\hat{t})$ is a residual set, because the countable intersection of residual sets is a residual set. The application of Lemma 1 completes the proof. □

References:

[1] J. L. Brown, Jr. On the error in reconstructing a non-bandlimited function by means of the bandpass sampling theorem. *Journal of Mathematical Analysis and Applications*, 18:75–84, 1967. Erratum, ibid, vol. 21, 1968, p. 699.

[2] P. L. Butzer, W. Splettstößer, and R. L. Stens. The sampling theorem and linear prediction in signal analysis. *Jahresber. d. Dt. Math.-Verein.*, 90(1):1–70, January 1988.

[3] P. L. Butzer and R. L. Stens. Sampling theory for not necessarily band-limited functions: A historical overview. *SIAM Review*, 34(1):40–53, March 1992.

[4] John R. Higgins. *Sampling Theory in Fourier and Signal Analysis – Foundations*. Oxford University Press, 1996.

[5] Kôaku Yosida. *Functional Analysis*. Springer-Verlag, 1971.

On Subordination Principles for Generalized Shannon Sampling Series

Andi Kivinukk [(1)] and Gert Tamberg [(2)]

(1) Dept. of Math., Tallinn University, Narva Road 25, 10120 Tallinn, Estonia
(2) Dept. of Math., Tallinn University of Technology, Ehitajate tee 5 19086 Tallinn, Estonia
andik@tlu.ee, gert.tamberg@mail.ee

Abstract:

This paper provides some subordination equalities and their applications for the generalized Shannon sampling series.

1. Introduction

For the uniformly continuous and bounded functions $f \in C(\mathbb{R})$ the generalized Shannon sampling series (see [3] and references cited there) are given by ($t \in \mathbb{R}$; $W > 0$)

$$(S_W f)(t) := \sum_{k=-\infty}^{\infty} f(\frac{k}{W}) s(Wt - k), \qquad (1)$$

where the condition for the operator $S_W : C(\mathbb{R}) \to C(\mathbb{R})$ to be well-defined is that for the kernel function $s = s(t)$ we assume

$$\sum_{k=-\infty}^{\infty} |s(u-k)| < \infty \quad (u \in \mathbb{R}).$$

Let be given an even window function $\lambda \in C_{[-1,1]}$, $\lambda(0) = 1$, $\lambda(u) = 0$ ($|u| \geqslant 1$,) then in our approach the kernel function will be defined by the equality

$$s(t) := s_\lambda(t) := \int_0^1 \lambda(u) \cos(\pi t u)\, du. \qquad (2)$$

Many window functions have been used in applications (see, e.g. [1], [2], [4], [8]), in Signal Analysis in particular. Next window functions are important for our subordination equalities.

1) $\lambda_{(r)}(u) = 1 - u^r$, $r \geq 1$ defines the Zygmund (or Riesz) kernel, denoted by $z_r = z_r(t)$, which special case $r = 1$, the Fejér (or Bartlett, see [8]) kernel $s_F(t) = \frac{1}{2}\mathrm{sinc}^2 \frac{t}{2}$, is well-known; the special case $r = 2$ is called also as the Welch [8] kernel;

2) $\lambda_j(u) := \cos \pi(j + 1/2)u$, $j = 0, 1, 2, \ldots$ defines the Rogosinski-type kernel (see [5]) in the form

$$r_j(t) := \frac{(-1)^j}{\pi} \frac{(j+1/2)\cos \pi t}{(j+1/2)^2 - t^2}; \qquad (3)$$

3) $\lambda_H(u) := \cos^2 \frac{\pi u}{2} = \frac{1}{2}(1 + \cos \pi u)$ defines the Hann kernel (see [6])

$$s_H(t) := \frac{1}{2} \frac{\mathrm{sinc} t}{1 - t^2}. \qquad (4)$$

Concerning some direct (Jackson-type) approximation theorems we present certain subordination equalities, which show that the sampling operators, like Rogosinski, Zygmund, and Hann, are in some sense basic.

2. Subordination equalities

Subordination equalities state some relations between two sampling operators.

2.1 Subordination by the Rogosinski-type sampling series

Let consider the Rogosinski-type sampling operators $R_{W,j}$ defined by the kernel functions r_j in (3). These kernel functions are deduced by the window functions $\lambda_j(u) := \cos \pi(j + 1/2)u$, $(j \in N)$ and as a family of functions it forms an orthogonal system on $[0, 1]$. Therefore, we may represent a quite arbitrary window function λ by its Fourier series. But the Fourier representation allows us to prove for a given kernel function s the sampling series

$$s(t) = 2 \sum_{j=0}^{\infty} s(j + 1/2)\, r_j(t).$$

In following B_σ^p stands for the Bernstein class, it consists of those bounded functions $f \in L^p(\mathbb{R})$ ($1 \leqslant p \leqslant \infty$), which can be extended to an entire function $f(z)$ ($z \in \mathbb{C}$) of exponential type σ. For $s \in B_\pi^1$ the sampling series above is absolutely convergence and by (1) we get formally the equalities

$$S_W f = 2 \sum_{j=0}^{\infty} s(j + 1/2) R_{W,j} f,$$

$$f - S_W f = 2 \sum_{j=0}^{\infty} s(j + 1/2)(f - R_{W,j} f),$$

calling as the subordination equalities, since the approximation properties of the general sampling operators (1) can be described via the approximation properties of the Rogosinski-type sampling operators $R_{W,j} : C(\mathbb{R}) \to C(\mathbb{R})$. We have proved that [5]

$$\|R_{W,j}\| = \frac{4}{\pi} \sum_{\ell=0}^{2j} \frac{1}{2\ell + 1} = \frac{2}{\pi} \log(j + 1) + O(1),$$

thus the subordination equalities are valid, when

$$\sum_{j=0}^{\infty} |s(j+1/2)| \log(j+1) < \infty.$$

Similar subordination equalities can be deduced for some interpolating sampling series, i.e. for which the equation $(\tilde{S}_W f)(\frac{k}{W}) = f(\frac{k}{W}) \quad (k \in \mathbb{Z})$ is valid. In [7] we have proved that the interpolating sampling operators will be defined by (1) using the kernel $\tilde{s}(t) := 2s(2t)$, where the kernel s is generated by (2) with a window function λ for which $\lambda(u) + \lambda(1-u) = 1 \quad (u \in [0,1])$.
Let the operator $S_W^\alpha : C(\mathbb{R}) \to C(\mathbb{R})$ be defined by the kernel $s_\alpha := \alpha\, s(\alpha\cdot) \in B^1_{\alpha\pi} \quad (0 < \alpha \le 2)$, where $s \in B^1_\pi$, and the modified Hann operator $H^\alpha_{W,j}$ is defined by the kernel

$$s^\alpha_{H,j}(t) := \frac{\alpha}{2}\, \frac{(2j+1)^2}{(2j+1)^2 - (\alpha t)^2} \operatorname{sinc}(\alpha t). \qquad (5)$$

Then here we have (see [7], Th. 2.3 and 2.4)

$$S^\alpha_W f = 4 \sum_{j=0}^{\infty} s(2j+1) H^\alpha_{W,j} f,$$

$$f - S^\alpha_W f = 4 \sum_{j=0}^{\infty} s(2j+1)(f - H^\alpha_{W,j} f).$$

2.2 Subordination by the Rogosinski-type sampling series: 2D case

The two-dimensional generalized sampling series has the form

$$(S_W f)(x,y) := \sum_{k,l=-\infty}^{\infty} f(\frac{k}{W}, \frac{l}{W}) s(Wx-k, Wy-l),$$

in particular, the multiplicative Rogosinski-type sampling series we define as

$$(R_{W;i,j} f)(x,y) := \sum_{k,l=-\infty}^{\infty} f(\frac{k}{W}, \frac{l}{W}) r_i(Wx-k) r_j(Wy-l),$$

where the Rogosinski-type kernel r_j is defined by (3). Here our subordination equalities read as

$$S_W f = 4 \sum_{i,j=0}^{\infty} s(i+1/2, j+1/2) R_{W;i,j} f,$$

$$f - S_W f = 4 \sum_{i,j=0}^{\infty} s(i+1/2, j+1/2)(f - R_{W;i,j} f),$$

provided

$$\sum_{i,j=1}^{\infty} |s(i+1/2, j+1/2)| \log i \log j < \infty.$$

By given subordination equalities we see that the non-multiplicative sampling series may be studied by the multiplicative Rogosinski-type sampling series.

2.3 Subordination by the Zygmund sampling series

The Zygmund sampling operator Z^r_W will be defined by the window function $\lambda_{(r)}(u) = 1 - u^r, \ r \ge 1$. Let us consider the kernel s in (2), for which the corresponding window function has the power series representation

$$\lambda(u) = 1 - \sum_{j=r}^{\infty} c_j u^j.$$

Then the formal subordination equalities are in the shape

$$S_W f = \sum_{j=r}^{\infty} c_j Z^j_W f,$$

$$f - S_W f = \sum_{j=r}^{\infty} c_j (f - Z^j_W f).$$

Several other subordination equalities and their applications will be presented.

3. Acknowledgments

This research was partially supported by the Estonian Sci. Foundation, grants 6943, 7033, and by the Estonian Min. of Educ. and Research, projects SF0132723s06, SF0140011s09.

References:

[1] H. H. Albrecht. A family of cosine-sum windows for high resolution measurements. In *IEEE International Conference on Acoustics, Speech and Signal Processing, Salt Lake City, Mai 2001*, pages 3081–3084. Salt Lake City, 2001.

[2] R. B. Blackman and J. W. Tukey. *The measurement of power spectra*. Wiley-VCH, New York, 1958.

[3] P. L. Butzer, G. Schmeisser, and R. L. Stens. An introduction to sampling analysis. In F Marvasti, editor, *Nonuniform Sampling, Theory and Practice*, pages 17–121. Kluwer, New York, 2001.

[4] F. J. Harris. On the use of windows for harmonic analysis. *Proc. of the IEEE*, 66:51–83, 1978.

[5] A. Kivinukk and G. Tamberg. On sampling series based on some combinations of sinc functions. *Proc. of the Estonian Academy of Sciences. Physics Mathematics*, 51:203–220, 2002.

[6] A. Kivinukk and G. Tamberg. On sampling operators defined by the Hann window and some of their extensions. *Sampling Theory in Signal and Image Processing*, 2:235–258, 2003.

[7] A. Kivinukk and G. Tamberg. Interpolating generalized Shannon sampling operators, their norms and approximation properties. *Sampling Theory in Signal and Image Processing*, 8:77–95, 2009.

[8] E. H. W. Meijering, W. J. Niessen, and M. A. Viergever. Quantitative evaluation of convolution-based methods for medical image interpolation. *Medical Image Analysis*, 5:111–126, 2001.

Linear Signal Reconstruction from Jittered Sampling

Alessandro Nordio [(1)], Carla-Fabiana Chiasserini [(1)] and Emanuele Viterbo [(2)]

(1) Dipartimento di Elettronica, Politecnico di Torino[1], I-10129 Torino, Italy.
(2) DEIS, Università della Calabria, via P. Bucci, Cubo 42C, 87036 Rende (CS), Italy
alessandro.nordio@polito.it, carla.chiasserini@polito.it, viterbo@deis.unical.it

Abstract:

This paper presents an accurate and simple method to evaluate the performance of AD/DA converters affected by clock jitter, which is based on the analysis of the mean square error (MSE) between the reconstructed signal and the original one. Using an approximation of the linear minimum MSE (LMMSE) filter as reconstruction technique, we derive analytic expressions of the MSE. Through asymptotic analysis, we evaluate the performance of digital signal reconstruction as a function of the clock jitter, number of quantization bits, signal bandwidth and sampling rate.

1. Introduction

A significant problem in Analog Digital Conversion (ADC) of wide-band signals is clock jitter and its impact on the quality of signal reconstruction. Indeed, even small amounts of jitter can measurably degrade the performance of analog to digital and digital to analog converters.
Clock jitter is typically detrimental because the analog to digital process relies upon a sample clock to indicate when a sample or snapshot of the analog signal is taken. The sample clock must be evenly spaced in time; any deviation will result in a distortion of the digitization process. If one had a perfect ADC and a perfect DAC and used the same clock to drive both units, then jitter would not have any impact on the reconstructed signal. In a real world system, however, a digitized signal travels through multiple processors, usually it is stored on a disk or piece of tape for a while, and then goes through more processing before being converted back to analog. Thus, during reconstruction, the clock pulses used to sample the signal are replaced with newer ones with their own subtle variations. Jitter may have different probability distributions which may have different effects on the quality of the reconstructed signal.
While several results are available in the literature on jittered sampling [4, 5] as well as on experimental measurements and instruments performance [1, 3, 6, 7], an analytical methodology for the performance study of the AD/DA conversion is still missing.
In this paper we fill this gap and propose a method for evaluating the performance of AD/DC converters affected by jitter, which is based on the analysis of the mean square error (MSE) between the reconstructed signal and the original one [7].

As reconstruction technique, we consider linear filtering methods, which typically have low complexity and are used in a wide variety of fields. If jitter were known exactly, the linear minimum MSE (LMMSE) reconstruction technique would be optimal, since it minimizes the MSE of the reconstructed signal. In practice this is not the case, hence we apply a reconstruction filter with the same structure of the LMMSE filter, where we let the jitter vanish. Then, we apply asymptotic analysis to derive analytical expressions of the MSE on the quality of the reconstructed signal. We then show that our asymptotic expressions provide an excellent approximation of the MSE even for small values of the system parameters, with the advantage of greatly reducing the computation complexity. We apply our method to study the performance of the AD/DA conversion system as a function of the clock jitter, number of quantization bits, signal bandwidth and sampling rate.

2. System model

Throughout the paper we use the following notations. Column vectors are denoted by bold lowercase letters and matrices are denoted by bold upper case letters. The (k, q)-th entry of the generic matrix $\mathbf{Z}$ is denoted by $(\mathbf{Z})_{k,q}$. The $n \times n$ identity matrix is denoted by $\mathbf{I}_n$, while $\mathbf{I}$ is the generic identity matrix. $(\cdot)^{\mathrm{T}}$ is the transpose operator, while $(\cdot)^{\dagger}$ is the conjugate transpose operator. We denote by $f_x(z)$ the probability density function (pdf) of the generic random variable x, and by $\mathbb{E}[\cdot]$ the average operator.

2.1 Signal sampling and reconstruction

We consider an analog signal $s(t)$ sampled at constant rate $f_s = 1/T_s$ over the finite interval $[0, MT_s)$. T_s is the sample spacing. When observed over a finite interval, $s(t)$ admits an infinite Fourier series expansion. Let N' denote the largest index of the non-negligible Fourier coefficients, then N'/T_s can be considered as the approximate one-sided bandwidth of the signal. We therefore represent the signal by using a truncated Fourier series with $N = 2N' +$

This work was supported by Regione Piemonte through the VICSUM project.

1 complex harmonics as

$$s(t) = \frac{1}{\sqrt{N}} \sum_{\ell=-N'}^{N'} a_\ell \exp\left(\mathrm{j}2\pi\ell \frac{t}{MT_s}\right), \qquad (1)$$

$0 \le t < MT_s$. The vector $\mathbf{a} = [a_{-N'}, \ldots, a_0, \ldots, a_{N'}]^{\mathrm{T}}$ represents the complex discrete spectrum of the signal.
Observe that the signal representation given in (1) includes sine waves of any fractional frequency $f_0 = f_s N'/M$ (when $a_\ell = 0$ for $-N' < \ell < N'$ and $a_{-N'} = a^*_{N'}$), which are frequently used as reference signal for calibration of ADC [1, 2]. We note that when the signal $s(t)$ is observed in the frequency domain through its M samples, the spectral resolution is given by $\Delta f = 1/(MT_s)$. Therefore, considering the expression in (1), the signal bandwidth is given by $B = \frac{N\Delta f}{2} = \frac{N}{2MT_s}$. By defining the parameter

$$\beta = \frac{M}{N} \qquad (2)$$

as the *oversampling factor* of the signal $s(t)$ with respect to the Nyquist rate, we can also write:

$$B = \frac{f_s/2}{\beta} \qquad (3)$$

In this work, we consider that sampling locations suffer from jitter, i.e., the m-th sampling location is given by

$$t_m = mT_s + d_m, \qquad (4)$$

$m = 0, \ldots, M-1$, where d_m is the associated independent random jitter whose distribution is denoted by $f_d(z)$. Typically, we have $|d_m| \ll T_s$.
Let the signal samples be $\mathbf{s} = [s_0, \ldots, s_{M-1}]^{\mathrm{T}}$ where $s_m = s(t_m)$, $0 \le m \le M-1$. Using (1), the set of signal samples can be written as

$$\mathbf{s} = \mathbf{V}^\dagger \mathbf{a}$$

where $\mathbf{V}$ is an $N \times M$ random Vandermonde matrix defined as

$$(\mathbf{V})_{\ell,m} = \frac{1}{\sqrt{N}} \exp\left(-\mathrm{j}2\pi\ell \frac{t_m}{MT_s}\right) \qquad (5)$$

$\ell = -N', \ldots, N'$, and $m = 0, \ldots, N-1$. Note that $\mathbf{V}$ accounts for the jitter in the AD/DA conversion process, and that the parameter β defined in (2) also represents the *aspect ratio* of matrix $\mathbf{V}$.
Furthermore, in addition to jittered sampling, we assume that signal samples are affected by some additive noise and are therefore given by

$$\mathbf{y} = \mathbf{s} + \mathbf{n}$$

where $\mathbf{n}$ is a vector of M noise samples, modeled as zero mean i.i.d. random variables. In practice, the dominant additive noise error is due to the n-bit quantization process [10].
In order to reconstruct the signal we consider a reconstruction technique that provides an estimate $\hat{\mathbf{a}}$ of the discrete spectrum $\mathbf{a}$. The reconstruction $\hat{s}(t)$ of $s(t)$ obtained from $\hat{\mathbf{a}}$ is given by

$$\hat{s}(t) = \frac{1}{\sqrt{N}} \sum_{\ell=-N'}^{N'} \hat{a}_\ell \exp\left(\mathrm{j}2\pi\ell \frac{t}{MT_s}\right)$$

2.2 Reconstruction error

We consider as performance metric of the AD/DA conversion process the mean square error (MSE) associated to the estimate. The MSE, evaluated in the observation interval $[0, MT_s)$, can be equivalently computed in both time and frequency domains as:

$$\mathrm{MSE} = \mathbb{E}\left[\int_0^{MT_s} |s(t) - \hat{s}(t)|^2 \,\mathrm{d}t\right] = \frac{\mathbb{E}\left[\|\mathbf{a} - \hat{\mathbf{a}}\|^2\right]}{N}$$

More specifically, we consider the MSE relative to the signal average power, i.e.,

$$J = \frac{\mathrm{MSE}}{\sigma_a^2}$$

which can be thought of as a noise to signal ratio and will be plotted using a dB scale in our results.
Among the possible techniques that can be applied to reconstruct the original signal, we focus on linear filters that provide an estimate of $\mathbf{a}$ through the linear operation $\hat{\mathbf{a}} = \mathbf{B}\mathbf{y}$ where $\mathbf{B}$ is an $N \times M$ matrix.

3. Jittered AD/DA conversion with linear filtering

Let us assume $\|\mathbf{a}\|^2 = \sigma_a^2 N$ and $\mathbb{E}[\mathbf{n}\mathbf{n}^\dagger] = \sigma_n^2 \mathbf{I}$, then we define the signal to noise ratio (SNR) in absence of jitter as

$$\gamma = \frac{\sigma_a^2}{\sigma_n^2}$$

Under the assumption that $\mathbb{E}[\mathbf{a}\mathbf{a}^\dagger] = \sigma_a^2 \mathbf{I}$, the linear filter that provides the best performance in terms of MSE is the linear minimum mean square error (LMMSE) filter, which is given by

$$\mathbf{B}_{\mathrm{opt}} = \left(\mathbf{V}\mathbf{V}^\dagger + \frac{1}{\gamma}\mathbf{I}\right)^{-1} \mathbf{V} \qquad (6)$$

In [8], it has been shown that, by applying the LMMSE filter, we obtain:

$$J = \frac{1}{\sigma_a^2 N} \mathbb{E}\left[\|\mathbf{a} - \hat{\mathbf{a}}\|^2\right] = \mathbb{E}\left[\mathrm{tr}\left\{\left(\gamma\mathbf{V}\mathbf{V}^\dagger + \mathbf{I}\right)^{-1}\right\}\right]$$

where $\mathrm{tr}\{\cdot\}$ is the normalized matrix trace operator and the average is over the randomness in $\mathbf{V}$.
Note, however, that the filter in (6) cannot be employed in practice, since the jitters d_m (hence the matrix $\mathbf{V}$) are unknown (see the definition of $\mathbf{V}$ in (5)). We therefore resort to an approximation of the optimum filter $\mathbf{B}_{\mathrm{opt}}$, based on the assumption that jitter has a zero mean.
In particular, we approximate $\mathbf{V}$ with the matrix $\mathbf{F}$ defined as,

$$\mathbf{F} = \mathbf{V}|_{d_m=0}$$

with the generic element of $\mathbf{F}$ given by, $(\mathbf{F})_{\ell,m} = \exp\left(-\mathrm{j}2\pi\ell\frac{m}{M}\right)/\sqrt{N}$, $\ell = -N', \ldots, N'$, and $m = 0, \ldots, N-1$. We observe that $\mathbf{F}$ is such that: $\mathbf{F}\mathbf{F}^\dagger = \beta\mathbf{I}$ and it is related to the discrete Fourier transform matrix.
Substituting the approximation of $\mathbf{V}$ in (6), we obtain:

$$\mathbf{B} = \left(\beta + \frac{1}{\gamma}\right)^{-1} \mathbf{F} \qquad (7)$$

Notice that the filter in (7) is the LMMSE filter adapted to the linear model $\mathbf{y} = \mathbf{F}^\dagger \mathbf{a} + \mathbf{n}$. By letting $\omega = (\beta + 1/\gamma)^{-1}$, the noise to signal ratio J provided by the approximate filter (7) is given by

$$\begin{aligned} J &= \frac{1}{\sigma_a^2 N} \mathbb{E}\left[\|\mathbf{a} - \omega \mathbf{F}\mathbf{y}\|^2\right] \\ &= \mathsf{tr}\left\{\omega^2 \mathop{\mathbb{E}}_{d}\left[\mathbf{F}\mathbf{V}^\dagger \mathbf{V}\mathbf{F}^\dagger\right] - 2\omega \Re\{\mathop{\mathbb{E}}_{d}\left[\mathbf{F}\mathbf{V}^\dagger\right]\}\right\} \\ &\quad + 1 + \frac{\omega^2 \beta}{\gamma} \end{aligned} \tag{8}$$

where the operator $\mathop{\mathbb{E}}_{d}[\cdot]$ averages over the random jitters d_m, $m = 0, \ldots, M-1$.
Assuming the jitters to be independent [1] and with characteristic function $C_d(w) = \mathop{\mathbb{E}}_{d}[\exp(\mathrm{j}wz)]$, the first two terms in (8) are given by

$$\mathsf{tr}\mathop{\mathbb{E}}_{d}\left[\mathbf{F}\mathbf{V}^\dagger\right] = \frac{\beta}{N} \sum_{\ell=-N'}^{N'} C_d\left(\frac{2\pi\ell}{MT_s}\right)$$

$$\mathop{\mathbb{E}}_{d}\left[\mathbf{F}\mathbf{V}^\dagger \mathbf{V}\mathbf{F}^\dagger\right] = \beta + \beta\frac{(\beta-1)}{N} \sum_{\ell=-N'}^{N'} \left|C_d\left(\frac{2\pi\ell}{MT_s}\right)\right|^2$$

Hence, we can write:

$$\begin{aligned} J &= 1 + \omega^2\beta\left(1 + \frac{1}{\gamma}\right) - 2\omega\frac{\beta}{N} \sum_{\ell=-N'}^{N'} C_d\left(\frac{2\pi\ell}{MT_s}\right) \\ &\quad + \omega^2\beta\frac{(\beta-1)}{N} \sum_{\ell=-N'}^{N'} \left|C_d\left(\frac{2\pi\ell}{MT_s}\right)\right|^2 \end{aligned} \tag{9}$$

In order to reduce the complexity of the computation of the reconstruction error and provide simple but accurate analytical tools, in the next section we let the parameters N and M go to infinity, while the ratio $\beta = M/N$ is kept constant. We therefore derive an asymptotic expression of J, which we will show well approximates the expression in (9) even for small N and M.

4. Asymptotic analysis

When N and M grow to infinity while β is kept constant, we define the *asymptotic* noise to signal ratio J as:

$$J_\infty^{(\beta,\gamma)} = \lim_{\substack{N,M\to+\infty \\ \beta}} J$$

In [8], it has been shown that $J_\infty^{(\beta,\gamma)}$ provides an excellent approximation of MSE/σ_a^2 even for small values of N and M, with the advantage of greatly simplifying the computation.
In the limit $N, M \to \infty$ with constant β, we compute

$$\begin{aligned} \mu_1 &= \lim_{\substack{N,M\to+\infty \\ \beta}} \frac{1}{N} \sum_{\ell=-N'}^{N'} C_d\left(\frac{2\pi\ell}{MT_s}\right) \\ &= \int_{-1/2}^{1/2} C_d\left(4\pi B x\right)\, \mathrm{d}x \end{aligned} \tag{10}$$

where, from (3), we used the fact that $1/\beta T_s = f_s/\beta = 2B$. Similarly, we define

$$\begin{aligned} \mu_2 &= \lim_{\substack{N,M\to+\infty \\ \beta}} \frac{1}{N} \sum_{\ell=-N'}^{N'} C_d\left(\frac{2\pi\ell}{MT_s}\right)^2 \\ &= \int_{-1/2}^{1/2} \left|C_d\left(4\pi B x\right)\right|^2 \mathrm{d}x \end{aligned} \tag{11}$$

By using (10) (11), and (9), the asymptotic expression of J is given by

$$J_\infty^{(\beta,\gamma)} = 1 + \omega^2\beta(1+1/\gamma) - 2\omega\beta\mu_1 + \omega^2\beta(\beta-1)\mu_2 \tag{12}$$

It is worth mentioning that for large SNRs (i.e., in absence of measurement noise), $J_\infty^{(\beta,\gamma)}$ reduces to

$$J_\infty^{(\beta)} = \lim_{\gamma\to\infty} J_\infty^{(\beta,\gamma)} = 1 + \frac{1}{\beta} - 2\mu_1 + \left(1 - \frac{1}{\beta}\right)\mu_2 \tag{13}$$

Equation (13) provides us with a floor that represent the best quality of the reconstructed signal (minimum MSE) we can hope for.

4.1 Example: uniform jitter distribution

Let us now assume the jitter to be uniformly distributed with pdf given by

$$f_d(z) = \begin{cases} \frac{1}{2d_{\max}} & -d_{\max} \le z \le d_{\max} \\ 0 & \text{elsewhere} \end{cases}$$

where $d_{\max}$ is the maximum jitter, independent of the sampling frequency f_s. In this case, the characteristic function of the jitter is given by $C_d(w) = \sin(d_{\max}w)/(d_{\max}w)$. Then,

$$\mu_1 = \frac{\mathrm{Si}(2\pi\eta_u)}{2\pi\eta_u}$$

and

$$\mu_2 = \frac{\cos^2(2\pi\eta_u) + 2\pi\eta_u \mathrm{Si}(4\pi\eta_u) - 1}{4\pi^2\eta_u^2}$$

where $\mathrm{Si}(\cdot)$ is the integral sine function and $\eta_u = d_{\max}B$ is a dimensionless parameter which relates maximum jitter and signal bandwidth.

5. Results

For the ease of representation, we assume that the dominant component of the additive noise is due to quantization, and we express the SNR in absence of jitter, γ, as a function of the number of quantization bits n of the ADC [9]:

$$(\gamma)_{\mathrm{dB}} = 6.02n + 1.76$$

Then, in the following plots we show the value of J as a function of γ or, equivalently, of the number of quantization bits n.
Figure 1 compares the value of J obtained through its asymptotic expression against the performance of a system with finite parameters values (i.e., the value of J computed using (9)). The results are derived for $\eta_u =$

$10^{-1}, 10^{-2}, 10^{-3}$, and $\beta = 10$. Solid lines refer to the asymptotic expression (12), while markers represent the values of J computed through (9), with $N' = 100$. We observe an excellent matching between our approximation of $J_{\infty}^{(\beta,\gamma)}$ and the results computed through (9), even for small values of N and M. We point out that this tight match can be observed for any $\beta > 1$ and $\eta_u \ll 1$.

We also notice that J shows a floor, whose expression is given by (13). This floor is due to the mismatch between the filter $\mathbf{F}$ employed in the reconstruction and the matrix $\mathbf{V}$ characterizing the sampling system.

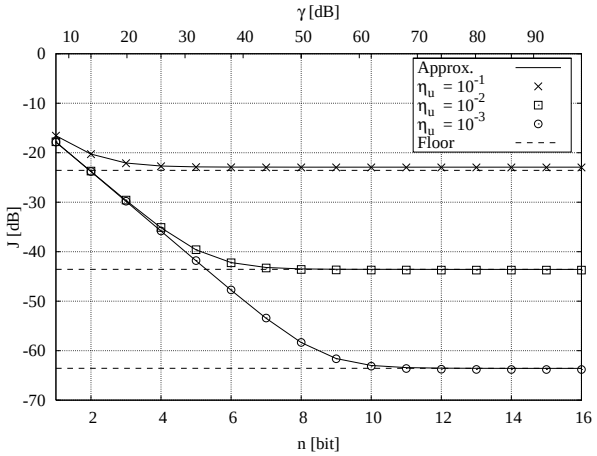


Figure 1: Comparison between the reconstruction error J derived through (9), the approximation of $J_{\infty}^{(\beta,\gamma)}$ and the floor $J_{\infty}^{(\beta)}$ in (13).

Furthermore, in the case of unknown jitter, and, thus, of a floor in the behavior of J, there exists a number of quantization bits $n = n^*$ beyond which a further increase in the ADC precision does not provide a noticeable decrease in the reconstruction error J. The relation between η_u, β, and n^* is shown in Figure 2. Note that n^* is lightly affected by an increase of β, provided that $\beta > 1$, and a good compromise for choosing the oversampling rate is $\beta = 5$.

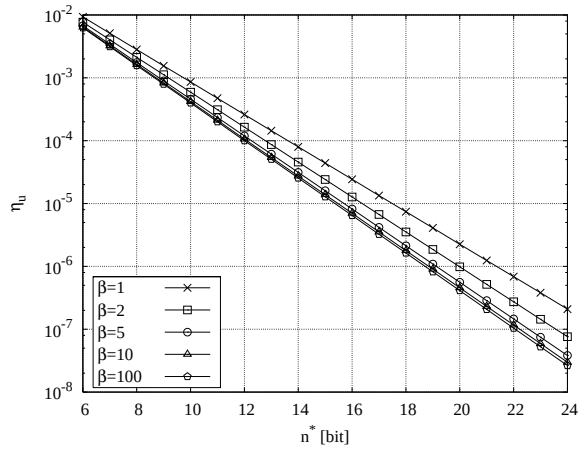


Figure 2: Minimum number of bits n^* required to reach the floor of $J_{\infty}^{(\beta,\gamma)}$ as a function of β and η_u.

6. Conclusions

We studied the performance of AD/DA converters, in presence of clock jitter and quantization errors. We considered that a linear filter approximating the LMMSE filter is used for signal reconstruction, and evaluated the system performance in terms of MSE. Through asymptotic analysis, we derived analytical expressions of the MSE which provide an accurate and simple method to evaluate the behavior of AD/DA converters as clock jitter, number of quantization bits, signal bandwidth and sampling rate vary. We showed that our asymptotic approach provides an excellent approximation of the MSE even for small values of the system parameters. Furthermore, we derived the MSE floor, which represents the best reconstruction quality level we can hope for and gives useful insights for the design of AD/DA converters.

References:

[1] *Project DYNAD, SMT4-CT98, Draft Standard Version 3.4*, Jul. 12, 2001.

[2] IEEE Standard for Terminology and Test Methods for Analog-to-Digital Converters, IEEE Std. 1241, 2000.

[3] P. Arpaia, P. Daponte, and S. Rapuano, "Characterization of digitizer timebase jitter by means of the Allan variance," *Computer Standards & Interfaces*, Vol. 25, pp. 15–22, 2003.

[4] B. Liu, and T. P. Stanley, "Error bounds for jittered sampling," *IEEE Transactions on Automatic Control*, Vol. 10, No. 4, pp. 449–454, Oct. 1965.

[5] J. Tourabaly, and A. Osseiran, "A jittered-sampling correction technique for ADCs," *IEEE International Workshop on Electronic Design, Test and Applications*, pp. 249–252, Los Alamitos, CA, USA, 2008.

[6] E. Rubiola, A. Del Casale, and A. De Marchi, "Noise induced time interval measurement biases," *46th IEEE Frequency Control Symposium*, pp. 265–269, May 1992.

[7] J. Verspecht, "Accurate spectral estimation based on measurements with a distorted-timebase digitizer," *IEEE Trans. on Instrumentation and Measurement* Vol. 43, pp. 210–215, Apr. 1994.

[8] A. Nordio, C.-F. Chiasserini, and E. Viterbo "Performance of linear field reconstruction techniques with noise and uncertain sensor locations," *IEEE Trans. on Signal Processing*, Vol. 56, No. 8, pp. 3535–3547, Aug. 2008.

[9] G. Gielen, "Analog building blocks for signal processing," ESAT-MICAS, Leuven, Belgium, 2006.

[10] S. C. Ergen, and P. Varaiya, "Effects of A-D conversion nonidealities on distributed sampling in dense sensor networks," IPSN '06, Nashville, Tennessee, Apr. 2006.

Uniform Sampling and Reconstruction of Trivariate Functions

Alireza Entezari

E301 CSE Building, University of Florida, Gainesville, FL, USA.
entezari@cise.ufl.edu

Abstract:

The Body Centered Cubic (BCC) and Face Centered Cubic (FCC) lattices have been known to outperform the commonly-used Cartesian sampling lattice due to their improved spectral sphere packing properties. However, the Cartesian lattice has been widely used for sampling of trivariate functions with applications in areas such as biomedical imaging, scientific data visualization and computer graphics. The widespread use of Cartesian lattice is partly due to the availability of tensor-product approach that readily extend the univariate reconstruction methods to trivariate setting. In this paper we report on recent advances on non-separable reconstruction algorithms, based on box splines, for reconstruction of data sampled on the BCC and FCC lattices. It turns out that these box spline reconstructions are *faster* than the corresponding tensor-product B-spline reconstructions on the Cartesian lattice. This suggests that not only the BCC and FCC lattices are more accurate sampling patterns, their respective reconstruction methods are also more computationally efficient than the tensor-product reconstructions – a fact which is contrary to the common assumption among practitioners.

1. Introduction

Sampling and reconstruction play a vital role in visualization and computer graphics. Various volume rendering algorithms rely on accurate reconstruction as a key step since the quality and fidelity of the rendered image heavily depends on reconstruction. In image processing reconstruction is used in resampling, resizing, conversion, and manipulation of sampled data.

In the realm of sampling, the term **regular** is often used to refer to the case that the sampling grid is uniform. Although there has been significant research, recently, in non-uniform sampling (e.g., sparse sampling, compressed sensing), the regular sampling is the most commonly-used sampling scheme in practice [21].

When it comes to sampling multivariate functions, the tensor-product of uniform sampling, which forms a Cartesian lattice, is almost always the choice. The simple structure of the Cartesian lattice and its separable nature allows one to readily apply a tensor-product paradigm to many problems in a multi-dimensional setting. The power of the dimensionality reduction will remain the major reason that the Cartesian lattice is the preferred tool in numerical algorithms. The other attraction of the Cartesian lattice is that it simply exists in *any* dimension and often tools and theory extend to problems in a higher dimensional setting in a trivial manner.

However, the Cartesian lattice has been known to be an inefficient lattice from the sampling-theoretic point of view. Miyakawa [12] and then Petersen and Middleton [16] were among the first people to discover the superiority of sphere-packing and sphere-covering lattices for sampling multivariate functions. In particular they have demonstrated that Cartesian lattice is very inefficient for sampling multivariate functions.

2. Optimal Sampling Lattices

When sampling a multivariate function with a lattice, generated by (integer linear combinations of the columns of) a **sampling matrix**, $\boldsymbol{M}$, the spectrum of the signal is contained in the **Brillouin zone**. Brillouin zone is the Voronoi cell of the **reciprocal** lattice. The reciprocal lattice to the lattice $\boldsymbol{M}$ is generated by the columns of the matrix $2\pi\boldsymbol{M}^{-\top}$. The multivariate version of the Nyquist frequency is the boundary of the Brillouin zone.

Without a priori knowledge when sampling multivariate functions, one often assumes that the underlying function has features possibly in all directions. Therefore, without knowledge about particular orientations of high-frequency features, we need to capture an *isotropic* spectrum during the sampling process. Therefore, the objective of optimal sampling is to maximize the isotropic content of the Brillouin zone. In other words, the sampling lattice whose Brillouin zone has the largest inscribing (hyper) sphere is the best sampling lattice. Therefore, the optimal sampling lattice in any dimension is the lattice whose reciprocal lattice allows for the densest packing of spheres.

In the bivariate setting the hexagonal lattice is the best sampling lattice since its reciprocal lattice, which happens to be the dual hexagonal lattice, allows for the best packing of 2-D with disks. When compared to the commonly-used Cartesian lattice with the same sampling density, the hexagonal lattice allows for about 14% more information to be captured in the spectrum of the underlying signal. This is illustrated in Figure 1 as the area of inscribing disc to the Brillouin zone of the hexagonal lattice (i.e., hexagon) is larger than the area of inscribing disc to the Brillouin zone of the Cartesian lattice (i.e., square), even

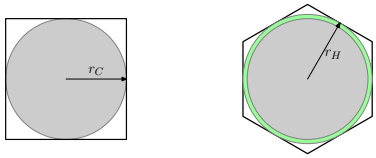


Figure 1: A square and a hexagon with unit area corresponding to the Brillouin zone of Cartesian and hexagonal sampling. The area of inscribing disk to a square is about 14% less than the area of the inscribing disk to the hexagon.

though the two Brillouin zones have the same area.

In the trivariate setting, the optimal sampling lattice is the BCC lattice whose reciprocal lattice (i.e., the FCC lattice) is the densest sphere packing lattice. The sampling efficiency of the BCC lattice, when compared to the commonly-used Cartesian lattice is about 30% higher. Appendix A in [6] presents a thorough comparison of the Brillouin zone of the Cartesian, BCC and FCC lattices.

The FCC lattice, is also superior to the Cartesian lattice as its efficiency compared to the Cartesian lattice is about 27% higher. Although among the FCC and BCC lattices the BCC wins, by a small margin, for optimal sampling, the FCC lattice appears to have good resistance to aliasing. This can be justified since its reciprocal lattice (i.e., the BCC lattice) allows for the best sphere covering of the space. The best covering of the space translates to replication of isotropic spectrum with minimal overlap between them– minimizing the aliasing for that sampling resolution.

These facts about comparison of the Cartesian, BCC and FCC lattices together with their higher-dimensional counter parts are discussed for sampling stationary isotropic random processes [10]. The arguments of the optimal sampling (BCC) and resilience to aliasing (FCC) is generalized to the notion that the reciprocal lattice for optimal sphere-packing lattice is the best choice for sampling functions at relatively high resolutions, while the sphere-packing lattice is the best option for sampling functions at relatively low resolutions [10].

3. Reconstruction

There is abundant research on reconstruction (i.e., interpolation or approximation) of data based on univariate filtering methods [15]. Various 1-D filters have a low-pass behavior and approximate the ideal kernel (i.e., sinc) for reconstruction into the space of band-limited functions. B-splines, offer a framework for representation of piecewise polynomial functions and thus are widely used in reconstruction of univariate functions [3].

There are two common methods for extending the univariate reconstruction 'kernels' to multivariate setting. The **separable** approach builds the multivariate kernel by a simple tensor-product of univariate kernels. The separable approach is obviously suitable for reconstruction of data on the Cartesian lattice since the lattice itself is also separable. The **radial** basis approaches construct the multivariate reconstruction kernel by spherical extension of univariate kernel. Due to the spherical extension, the radial basis approach ignores the underlying geometry of the sampling lattice and is often used for scattered data interpolation/approximation.

Splines have been widely accepted for image processing [20]. In the context of image processing, splines are often constructed as a tensor-product of two univariate splines. Mitchell and Netravali [11], demonstrated the advantages of using splines for image processing. Recently, Van De Ville [22], developed the so called Hex-splines that are used for reconstruction of hexagonal images. Hex-splines can not be constructed as a tensor-product of univariate splines. Due to the non-separable structure of hexagonal lattice, the tensor-product splines can not be applied for processing of hexagonal data.

3.1 Reconstruction of trivariate functions

In the visualization community reconstruction filters have received a lot of attention since accurate reconstruction of trivariate functions and their gradients is crucial in fidelity of rendering algorithms [14, 1, 5, 13]. Similar to image processing, in volume visualization algorithms, often the tensor-product approach is used for reconstruction of Cartesian sampled data.

Theußl [18] introduced the BCC sampling in volume rendering. However, since the BCC lattice is a non-separable lattice, various ad-hoc tensor-product [17] and radial basis [18] algorithms fail to provide satisfactory reconstruction algorithms and they exhibit blurry artifacts. Csébfalvi [2] proposed a global pre-processing algorithm (based on generalized interpolation [19]) that reconstructs the BCC lattice based on its two Cartesian sub-lattices. This approach is computationally inefficient and does not guarantee approximation order.

The author's recent work in this area establishes the relationship between box splines and the above-mentioned sampling lattices. The box splines have been developed as a generalization of B-splines to the multivariate setting. While box splines have been considered as non-separable basis functions for approximation based on their shifts on the Cartesian lattice [4], here their shifts on BCC and FCC lattices are considered. The interesting fact about these box splines is that while their shifts on the Cartesian lattice do not form a linearly independent set of functions, their shifts on the FCC and BCC lattices are linearly independent – a rare and useful property for the spline space!

3.2 Four direction box splines on BCC

The relation of box splines with the BCC lattice was established based on the fact that the immediate neighborhood of a lattice point on the BCC pattern forms a rhombic dodecahedron (see Figure 2). This polyhedron has the special property that is a projection of a four-dimensional hypercube (tesseract). This makes it a perfect match to be the support of a box spline since the geometric definition of box splines precisely amounts to projecting hypercubes (i.e., box) down to lower dimensional spaces. Generally, the class of polytopes that are the shadow of higher dimensional hypercubes are referred to as **zonotopes**. This linear box spline is defined by the four direction and is

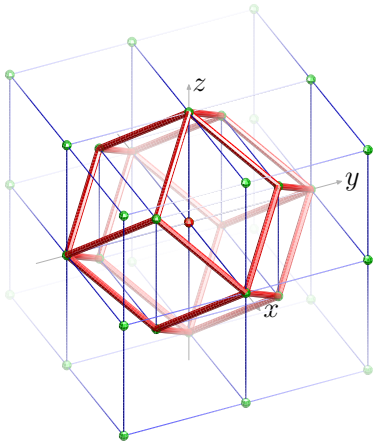


Figure 2: The neighborhood of a BCC lattice point forms a rhombic dodecahedron. This polyhedron is a zonohedron which is the support of a linear box spline.

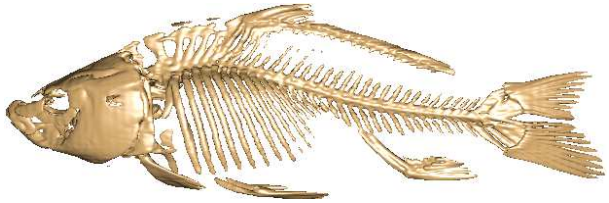


Figure 3: Benchmark example dataset. The CT dataset of a carp fish at a high resolution of $256 \times 256 \times 256$.

a C^0 kernel. The shifts of this box spline on the BCC lattice generate a spline space whose approximation order is two. By convolving this box spline by itself, one obtains a smoother, C^2, quintic box spline that is specified by a repetition of the four principal directions. The shifts of this box spline generate a spline space whose approximation order is four [7, 8]. This smoothness and approximation order match that of the tricubic B-spline on the Cartesian lattice and hence we compare the two on a Carp fish dataset in first row in Figure 4. The piecewise polynomial representation of these box splines along with efficient evaluation methods can be found in [8].

3.3 The six direction box spline on FCC

Unlike the BCC lattice, the immediate neighborhood in the FCC lattice is not a zonohedron. However, by enlarging the neighborhood one finds the truncated octahedron which is a zonohedron Figure 5. This polyhedron is a projection of a six-dimensional hypercube and the corresponding box spline is a cubic six-direction box spline [6]. The spline space that is generated by shifts of this cubic box spline on the FCC lattice is a C^1 space whose approximation order is three. These characteristics match the triquadratic B-spline on the Cartesian lattice which is the base for our comparisons in second row in Figure 4. The piecewise polynomial representation of the cubic box spline along with efficient spline evaluation method on the FCC lattice is demonstrated in [9].

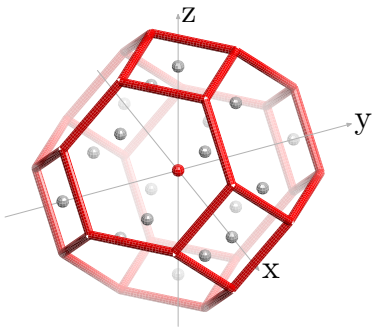


Figure 5: The neighborhood of a FCC lattice point forms a truncated octahedron. This polyhedron is another zonohedron which is the support of a six-direction box spline.

3.4 Computational advantages

Once efficient evaluation algorithms are derived for the four-direction box splines [8] and the six direction box spline [9], one can compare these box spline reconstructions to the commonly-used tensor-product B-spline reconstructions on the Cartesian lattice.
For the C^2, fourth-order method the tricubic B-spline uses a neighborhood of $4 \times 4 \times 4 = 64$ points for reconstruction, while the quintic box spline only uses a total of 32 points for reconstruction. Therefore as documented in [8] the BCC non-separable box spline approach outperforms the comparable tensor-product B-spline approach by a factor of **two**. Similarly the triquadratic B-spline uses a neighborhood of $3 \times 3 \times 3 = 27$ Cartesian data points, while the cubic box spline only requires a total of 16 FCC data points for the reconstruction. Therefore, the non-separable box spline reconstruction outperforms the comparable tensor-product B-spline approach as documented in [9].

4. Conclusions

The recent research on optimal sampling lattices suggests that not only the FCC and BCC lattices offer higher-fidelity sampling schemes, but also their reconstruction algorithms outperform the corresponding tensor-product reconstructions on the traditionally-popular Cartesian lattice. These encouraging results are crucial for acceptance of these efficient lattices in practical applications.

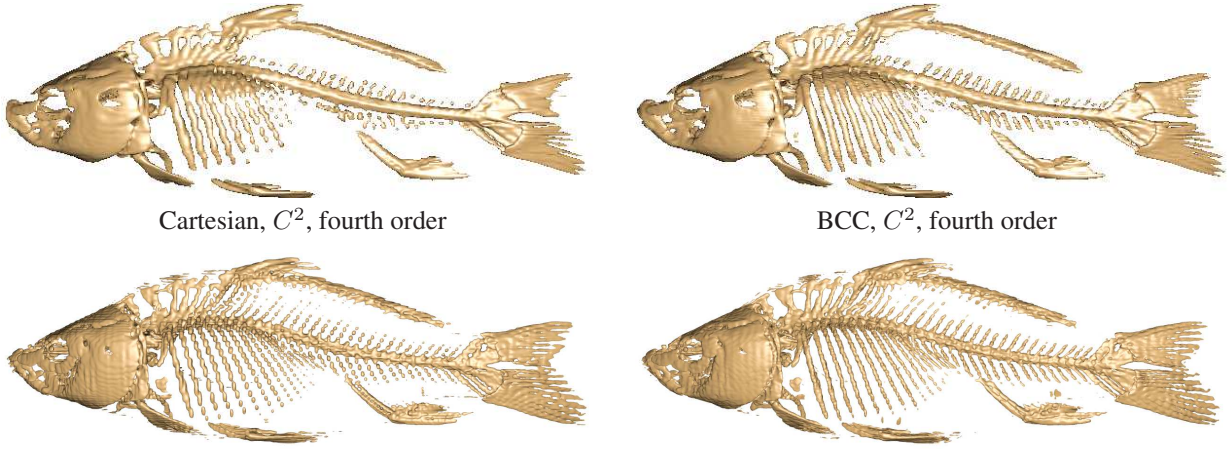
5. Acknowledgments

The author would like to thank Dimitri Van De Ville, Torsten Möller and Carl de Boor for valuable insight and advice at various stages of the work.

References:

[1] I. Carlbom. Optimal Filter Design for Volume Reconstruction and Visualization. In *Proc. IEEE Conf on Visualization*, pages 54–61, October 1993.

[2] B. Csébfalvi. Prefiltered gaussian reconstruction for high-quality rendering of volumetric data sampled



Cartesian, C^2, fourth order | BCC, C^2, fourth order

Cartesian, C^1, third order | FCC, C^1, third order

Figure 4: The Carp dataset at 6% resolution on Cartesian, BCC and FCC subsampled from the ground truth volume data of Figure 3. Top row: the Cartesian dataset is reconstructed by the tricubic B-spline and the BCC dataset is reconstructed by the quintic box spline. Bottom row: the Cartesian dataset is reconstructed with the triquadratic B-spline, while the FCC dataset is reconstructed with the cubic box spline. Superiority of the FCC and the BCC sampling is demonstrated since their images offer more accurate reconstruction than the Cartesian specially on the ribs and tail area.

on a body-centered cubic grid. In *IEEE Visualization*, pages 311–318, 2005.

[3] C. de Boor. *A practical guide to splines*, volume 27 of *Applied Mathematical Sciences*. Springer-Verlag, New York, revised edition, 2001.

[4] C. de Boor, K. Höllig, and S. Riemenschneider. *Box Splines*. Springer Verlag, 1993.

[5] S. C. Dutta Roy and B. Kumar. *Handbook of Statistics*, volume 10, chapter Digital Differentiators, pages 159–205. Elsevier Science Publishers B. V., N. Holland, 1993.

[6] A. Entezari. *Optimal Sampling Lattices and Trivariate Box Splines*. PhD thesis, Simon Fraser University, Vancouver, Canada, July 2007.

[7] A. Entezari, R. Dyer, and T. Möller. Linear and Cubic Box Splines for the Body Centered Cubic Lattice. In *Proceedings of the IEEE Conference on Visualization*, pages 11–18, October 2004.

[8] A. Entezari, D. Van De Ville, and T. Möller. Practical box splines for volume rendering on the body centered cubic lattice. *IEEE Trans. on Visualization and Comp Graphics*, 14(2):313 – 328, 2008.

[9] M. Kim, A. Entezari, and J. Peters. Box Spline Reconstruction on the Face Centered Cubic Lattice. *IEEE Trans. on Visualization and Computer Graphics*, 14(6):1523–1530, 2008.

[10] HR Kunsch, E. Agrell, and FA Hamprecht. Optimal lattices for sampling. *Information Theory, IEEE Transactions on*, 51(2):634–647, 2005.

[11] D. P. Mitchell and A. N. Netravali. Reconstruction Filters in Computer Graphics. In *Computer Graphics (Proceedings of SIGGRAPH 88)*, volume 22, pages 221–228, August 1988.

[12] H. Miyakawa. Sampling theorem of stationary stochastic variables in multidimensional space. *Journal of the Institute of Electronic and Communication Engineers of Japan*, 42:421–427, 1959.

[13] T. Möller, R. Machiraju, K. Mueller, and R. Yagel. A Comparison of Normal Estimation Schemes. In *Proceedings of the IEEE Conference on Visualization*, pages 19–26, October 1997.

[14] T. Möller, K. Mueller, Y. Kurzion, R. Machiraju, and R. Yagel. Design of Accurate and Smooth Filters for Function and Derivative Reconstruction. *Proceedings of the Symposium on Volume Visualization*, pages 143–151, Oct 1998.

[15] A.V. Oppenheim and R.W. Schafer. *Discrete-Time Signal Processing*. Prentice Hall Inc., Englewoods Cliffs, NJ, 1989.

[16] D. P. Petersen and D. Middleton. Sampling and Reconstruction of Wave-Number-Limited Functions in N-Dimensional Euclidean Spaces. *Information and Control*, 5(4):279–323, December 1962.

[17] T. Theußl, O. Mattausch, T. Möller, and E. Gröller. Reconstruction schemes for high quality raycasting of the body-centered cubic grid. *TR-186-2-02-11, Institute of Computer Graphics and Algorithms, Vienna University of Technology*, December 2002.

[18] T. Theußl, T. Möller, and E. Gröller. Optimal Regular Volume Sampling. In *Proc of the IEEE Conf on Visualization*, pages 91–98, Oct 2001.

[19] P. Thévenaz, T. Blu, and M. Unser. Interpolation revisited. *IEEE Transactions on Medical Imaging*, 19(7):739–758, July 2000.

[20] M. Unser. Splines: A perfect fit for signal and image processing. *IEEE Signal Processing Magazine*, 16(6):22–38, November 1999. IEEE Signal Processing Society's 2000 magazine award.

[21] M. Unser. Sampling—50 Years after Shannon. *Proceedings of the IEEE*, 88(4):569–587, April 2000.

[22] D. Van De Ville, T. Blu, M. Unser, W. Philips, I. Lemahieu, and R. Van de Walle. Hex-Splines: A Novel Spline Family for Hexagonal Lattices. *IEEE Trans. on Img Proc.*, 13(6):758–772, June 2004.

An Efficient Algorithm for the Discrete Gabor Transform using full length Windows

Peter L. Søndergaard

***Abstract*—This paper extends the efficient factorization of the Gabor frame operator developed by Strohmer in [17] to the Gabor analysis/synthesis operator. The factorization provides a fast method for computing the discrete Gabor transform (DGT) and several algorithms associated with it. The factorization algorithm should be used when the involved window and signal have the same length. An optimized implementation of the algorithm is freely available for download.**

I. Introduction

The finite, discrete Gabor transform (DGT) of a signal f of length L is given by

$$c\left(m,n,w\right)=\sum_{l=0}^{L-1}f(l,w)\overline{g\left(l-an\right)}e^{-2\pi iml/M}. \qquad (1)$$

Here g is a window (filter prototype) that localizes the signal in time and in frequency. The DGT is equivalent to a Fourier modulated filter bank with M channels and decimation in time a, [2].

Efficient computation of a DGT can be done by several methods: If the window g has short support (consists of relatively few filter taps), a filter bank based approach can be used. We shall instead focus on the case when g and f are equally long. The main advantage of the algorithm presented is its ease of use: The running time is guaranteed to be small even for long windows. This allows for the practical use of non-compactly supported windows like the Gaussian and its tight and dual windows without truncating them.

In the case when the window and signal have the same length, a factorization of the frame operator matrix was found by Zibulski and Zeevi in [19]. The method was initially developed in the $L^2\left(\mathbb{R}\right)$ setting, and was adapted for the finite, discrete setting by Bastiaans and Geilen in [1]. They extended it to also cover the analysis/synthesis operator. A simple, but not so efficient, method was developed for the Gabor analysis/synthesis operator by Prinz in [15]. Strohmer [17] improved the method and obtained the lowest known computational complexity for computing the Gabor frame operator. This paper extends Strohmer's method to also cover the Gabor analysis and synthesis operators.

The advantage of the method developed in this paper as compared to the one developed in [1], is that it works with FFTs of shorter length, and does not require multiplication by complex exponentials caused by the quasi-periodicity of the Zak transform. The two methods have the same asymptotic complexity, $O\left(NM\log M\right)$, where M is the number of channels and N is the number of time steps. A more accurate flop count is presented later in the paper.

We shall study the DGT applied to multiple signals at once. This is for instance a common subroutine in computing a multidimensional DGT. The DGT defined by (1) works on a multi-signal $f\in\mathbb{C}^{L\times W}$, where $W\in\mathbb{N}$ is the number of signals.

II. Definitions

We shall denote the set of integers between zero and some number L by

$$\langle L\rangle=0,\ldots,L-1. \qquad (2)$$

The Discrete Fourier Transform (DFT) of a signal $f\in\mathbb{C}^L$ is defined by

$$\left(\mathcal{F}_Lf\right)\left(k\right) \;=\; \frac{1}{\sqrt{L}}\sum_{l=0}^{L-1}f(l)e^{-2\pi ikl/L}. \qquad (3)$$

We shall use the $\cdot$ notation in conjunction with the DFT to denote the variable over which the transform is to be applied. To denote all elements indexed by a variable we shall use the : notation. As an example, if $C\in\mathbb{C}^{M\times N}$ then $C_{:,1}$ is a $M\times 1$ column vector, $C_{1,:}$ is a $1\times N$ row vector and $C_{:,:}$ is the full matrix. This notation is commonly used in Matlab and FORTRAN programming and also in some prominent textbooks, [8].

The convolution $f*g$ of two functions $f,g\in\mathbb{C}^L$ and the involution f^* is given by

$$\left(f*g\right)\left(l\right) \;=\; \sum_{k=0}^{L-1}f\left(k\right)g\left(l-k\right),\quad l\in\langle L\rangle \qquad (4)$$

$$f^*\left(l\right) \;=\; \overline{f\left(-l\right)},\quad l\in\langle L\rangle. \qquad (5)$$

It is well known how convolution can be computed efficiently using the discrete Fourier transform. We shall use a variant of this result

$$\left(f*g^*\right)\left(l\right) \;=\; \sqrt{L}\mathcal{F}_L^{-1}\left(\left(\mathcal{F}_Lf\right)\left(\cdot\right)\overline{\left(\mathcal{F}_Lg\right)\left(\cdot\right)}\right)\left(l\right). \qquad (6)$$

The Poisson summation formula in the finite, discrete setting is given by

$$\mathcal{F}_M\left(\sum_{k=0}^{b-1}g(\cdot+kM)\right)\left(m\right) \;=\; \sqrt{b}\left(\mathcal{F}_Lg\right)\left(mb\right), \qquad (7)$$

where $g\in\mathbb{C}^L$, $L=Mb$ with $b,M\in\mathbb{N}$.

A family of vectors e_j, $j\in\langle J\rangle$ of length L is called a *frame* if constants $0<A\leq B$ exist such that

$$A\left\|f\right\|^2\leq\sum_{j=0}^{J-1}\left|\langle f,e_j\rangle\right|^2\leq B\left\|f\right\|^2,\quad\forall f\in\mathbb{C}^L. \qquad (8)$$

Algorithm 1 Window factorization

WFAC(g, a, M)

1) **for** $r = \langle c \rangle$ $k = \langle p \rangle$, $l = \langle q \rangle$
2) **for** $s = \langle d \rangle$
3) $tmp\,(s) \leftarrow$ $g\,(r + c \cdot (k \cdot q - l \cdot p + s \cdot p \cdot q \textbf{ mod } d \cdot p \cdot q))$
4) **end for**
5) $Phi\,(r, k, l, :) \leftarrow$DFT$(tmp)$
6) **end for**
7) **return** Phi

Algorithm 2 Discrete Gabor transform

DGT(f, g, a, M)

1) $Phi =$WFAC(g, a, M)
2) **for** $r = \langle c \rangle$
3) **for** $k = \langle p \rangle$, $l = \langle q \rangle$, $w = \langle W \rangle$
4) **for** $s = \langle d \rangle$
5) $tmp\,(s) \leftarrow$ $f\,(r + (k \cdot M + s \cdot p \cdot M - l \cdot h_a \cdot a \bmod \mathrm{L}), w)$
6) **end for**
7) $Psitmp\,(k, l + w \cdot q, \cdot) \leftarrow$DFT$(tmp)$
8) **end for**
9) **for** $s = \langle d \rangle$
10) $G \leftarrow Phi\,(:, :, r, s)$
11) $F \leftarrow Psitmp\,(:, :, s)$
12) $Ctmp\,(:, :, s) \leftarrow G^T \cdot F$
13) **end for**
14) **for** $u = \langle q \rangle$, $l = \langle q \rangle$, $w = \langle W \rangle$
15) $tmp \leftarrow$IDFT$(Ctmp\,(u, l + w \cdot q, :))$
16) **for** $s = \langle d \rangle$
17) $coef\,(r + l \cdot c, u + s \cdot q - l \cdot h_a \bmod N, w)$ $\leftarrow tmp\,(s)$
18) **end for**
19) **end for**
20) **end for**
21) **for** $n = \langle N \rangle$,$w = \langle W \rangle$
22) $coef\,(:, n, w) \leftarrow$DFT$(coef\,(:, n, w))$
23) **end for**
24) **return** $coef$

The constants A and B are called lower and upper frame bounds. If $A = B$, the frame is called *tight*. If $J > L$, the frame is redundant (oversampled). Finite- and infinite dimensional frames are described in [4].

A finite, discrete *Gabor system* (g, a, M) is a family of vectors $g_{m,n} \in \mathbb{C}^L$ of the following form

$$g_{m,n}\,(l) = e^{2\pi i l m/M} g\,(l - na)\,, \quad l \in \langle L \rangle \tag{9}$$

for $m \in \langle M \rangle$ and $n \in \langle N \rangle$ where $L = aN$ and $M/L \in \mathbb{N}$. A Gabor system that is also a frame is called a *Gabor frame*. The analysis operator $C_g : \mathbb{C}^L \mapsto \mathbb{C}^{M \times N}$ associated to a Gabor system (g, a, M) is the DGT given by given by (1). The Gabor synthesis operator $D_\gamma : \mathbb{C}^{M \times N} \mapsto \mathbb{C}^L$ associated to a Gabor system (γ, a, M) is given by

$$f\,(l) = \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} c\,(m, n)\, e^{2\pi i m l/M} \gamma\,(l - an)\,. \tag{10}$$

In (1), (9) and (10) it must hold that $L = Na = Mb$ for some $M, N \in \mathbb{N}$. Additionally, we define $c, d, p, q \in \mathbb{N}$ by

$$c = \gcd\,(a, M) \quad , \quad d = \gcd\,(b, N)\,, \tag{11}$$

$$p = \frac{a}{c} = \frac{b}{d} \quad , \quad q = \frac{M}{c} = \frac{N}{d}, \tag{12}$$

where GCD denotes the greatest common divisor of two natural numbers. With these numbers, the *redundancy* of the transform can be written as $L/\,(ab) = q/p$, where q/p is an irreducible fraction. It holds that $L = cdpq$. The Gabor *frame operator* $S_g : \mathbb{C}^L \mapsto \mathbb{C}^L$ of a Gabor frame (g, a, M) is given by the composition of the analysis and synthesis operators $S_g = D_g C_g$. The Gabor frame operator is important because it can be used to find the *canonical dual window* $g^d = S_g^{-1} g$ and the *canonical tight window* $g^t = S_g^{-1/2} g$ of a Gabor frame. The canonical dual window is important because D_{g^d} is a left inverse of C_g. This gives an easy way to construct an inverse transform of the DGT. Similarly, then D_{g^t} is a left inverse of C_{g^t}. For more information on Gabor systems and properties of the operators C, D and S see [9], [6], [7].

III. The algorithm

We wish to make an efficient calculation of all the coefficients of the DGT. Using (1) literally to compute all coefficients $c\,(m, n, w)$ would require $8MNLW$ flops.

To derive a faster DGT, one approach is to consider the analysis operator C_g as a matrix, and derive a faster algorithm through unitary matrix factorizations of this matrix. This is the approach taken by [17], [16]. Unfortunately, this approach tends to introduce many permutation matrices and Kronecker product matrices. Another approach is the one taken in [1] where the Zak transform is used. This approach has the downside that values outside the fundamental domain of the Zak transform require an additional step to compute. In this paper we have chosen to derive the algorithm by directly manipulating the sums in the definition of the DGT.

To find a more efficient algorithm than (1), the first step is to recognize that the summation and the modulation term in (1) can be expressed as a DFT:

$$c\,(m, n, w) = \sqrt{L} \mathcal{F}_L \left(f(\cdot, w) \overline{g\,(\cdot - an)} \right) (mb)\,. \tag{13}$$

We can improve on this because we do not need all the coefficients computed by the Fourier transform appearing in (13), only every b'th coefficient. Therefore, we can rewrite by the Poisson summation formula (7):

$$\begin{aligned} & c\,(m, n, w) \\ = \; & \sqrt{M} \mathcal{F}_M \left(\sum_{\tilde{m}=0}^{b-1} f(\cdot + \tilde{m}M, w) \overline{g\,(\cdot + \tilde{m}M - an)} \right) (m) \\ = \; & \left(\mathcal{F}_M K\,(\cdot, n, w) \right) (m)\,, \end{aligned} \tag{14}$$

where

$$K\,(j, n, w) = \sqrt{M} \sum_{\tilde{m}=0}^{b-1} f\,(j + \tilde{m}M, w)\, \overline{g}\,(j + \tilde{m}M - na)\,, \tag{15}$$

for $j \in \langle M \rangle$ and $n \in \langle N \rangle$. From (14) it can be seen that computing the DGT of a signal f can be done by computing K followed by DFTs along the first dimension of K.

To further lower the complexity of the algorithm, we wish to express the summation in (15) as a convolution.

We split j as $j = r + lc$ with $r \in \langle c \rangle$, $l \in \langle q \rangle$ and introduce $h_a, h_M \in \mathbb{Z}$ such that the following is satisfied:

$$c = h_M M - h_a a. \tag{16}$$

The two integers h_a, h_M can be found by the extended Euclid algorithm for computing the GCD of a and M.

Using (16) and the splitting of j we can express (15) as

$$\begin{aligned} & K(r+lc, n, w) \\ = \; & \sqrt{M} \sum_{\tilde{m}=0}^{b-1} f(r+lc+\tilde{m}M, w) \times \\ & \times \overline{g}(r + l(h_M M - h_a a) + \tilde{m}M - na) \qquad (17) \\ = \; & \sqrt{M} \sum_{\tilde{m}=0}^{b-1} f(r+lc+\tilde{m}M, w) \times \\ & \times \overline{g}(r + (\tilde{m} + lh_M) M - (n + lh_a) a) \qquad (18) \end{aligned}$$

We substitute $\tilde{m} + lh_M$ by $\tilde{m}$ and $n + lh_a$ by n and get

$$\begin{aligned} & K(r+lc, n-lh_a, w) \\ = \; & \sqrt{M} \sum_{\tilde{m}=0}^{b-1} f(r+lc+(\tilde{m}-lh_M)M, w) \times \\ & \times \overline{g}(r+\tilde{m}M - na) \qquad (19) \\ = \; & \sqrt{M} \sum_{\tilde{m}=0}^{b-1} f(r+\tilde{m}M + l(c-h_M M), w) \times \\ & \times \overline{g}(r+\tilde{m}M - na) \qquad (20) \end{aligned}$$

We split $\tilde{m} = k + \tilde{s}p$ with $k \in \langle p \rangle$ and $\tilde{s} \in \langle d \rangle$ and $n = u + sq$ with $u \in \langle q \rangle$ and $s \in \langle d \rangle$ and use that $M = cq$, $a = cp$ and $c - h_M M = -h_a a$:

$$\begin{aligned} & K(r+lc, u+sq-lh_a, w) \\ = \; & \sqrt{M} \sum_{k=0}^{p-1} \sum_{\tilde{s}=0}^{d-1} f(r+kM+\tilde{s}pM - lh_a a, w) \times \\ & \times \overline{g}(r+kM - ua + (\tilde{s}-s)pM) \qquad (21) \end{aligned}$$

After having expressed the variables j, $\tilde{m}$, n using the variables r, s, $\tilde{s}$, k, l, u we have now indexed f using $\tilde{s}$ and g using $(\tilde{s} - s)$. This means that we can view the summation over $\tilde{s}$ as a convolution, which can be efficiently computed using a discrete Fourier transform. Define

$$\Psi^f_{r,s}(k, l+wq) = \mathcal{F}_d f(r + kM + \cdot pM - lh_a a, w), \qquad (22)$$

$$\Phi^g_{r,s}(k, u) = \sqrt{M} \mathcal{F}_d g(r + kM + \cdot pM - ua), \qquad (23)$$

Using (6) we can now write (21) as

$$\begin{aligned} & K(r+lc, u+\tilde{s}q - lh_a, w) \\ = \; & \sqrt{d} \sum_{k=0}^{p-1} \mathcal{F}_d^{-1} \left(\Psi^f_{r,\cdot}(k, l+wq) \, \overline{\Phi^g_{r,\cdot}(k,u)} \right)(\tilde{s}) \qquad (24) \\ = \; & \sqrt{d} \mathcal{F}_d^{-1} \left(\sum_{k=0}^{p-1} \Psi^f_{r,\cdot}(k, l+wq) \, \overline{\Phi^g_{r,\cdot}(k,u)} \right)(\tilde{s}) \qquad (25) \end{aligned}$$

Table I
FLOP COUNTS

Alg.:	Flop count
Eq. (1)	$8MNL$
Eq. (14)	$8L\frac{L_g}{a} + 4NM\log_2(M)$
[1]	$L\left(8q + 1 + \frac{q}{p}\right) + 4L\left(1 + \frac{q}{p}\right)\log_2 N + 4MN\log_2(M)$
Alg. 2	$L(8q) + 4L\left(1 + \frac{q}{p}\right)\log_2 d + 4MN\log_2(M)$

Flop counts for 4 different way of computing the DGT: By the linear algebra definition (1), by the method based on Poisson summation (14), by the method of Bastiaans and Geilen from [1] and by Algorithm 2. The term L_g denotes the length of the window used so L_g/a is the overlapping factor of the window. Note for comparison that $\log_2 N = \log_2 d + \log_2 q$

If we consider $\Psi^f_{r,s}$ and $\Phi^g_{r,s}$ as matrices for each r and s, the sum over k in the last line can be written as matrix products. Algorithm 2 follows from this.

IV. RUNNING TIME

When computing the flop count of the algorithm, we will assume that a complex FFT of length M can be computed using $4M\log_2 M$ flops. A nice review of flop counts for FFT algorithms is presented in [14]. Table I shows the flop count for Algorithm 2 and compares it with the definition of the DGT (1), with the algorithm for short windows using Poisson summation (14) and with the algorithm published in [1]. The algorithm by Prinz presented in [15] has the same computational complexity as the Poisson summation algorithm. For simplicity we assume that both the window and signal are complex valued. In the common case when both f and g are real-valued, all the algorithms will see a 2 to 4 times speedup.

The flop count for definition (1) is that of a complex matrix multiplication. All the other algorithms share the $4MN\log_2 M$ term coming from the application of an FFT to each 'block' of coefficients and only differ in how the application of the window is performed. The Poisson summation algorithm is very fast for a small overlapping factor L_g/a, but turns into an $\mathcal{O}(L^2)$ algorithm for a full length window. In this case the other algorithms have an advantage. The term $L\left(8q + 1 + \frac{q}{p}\right)$ in the [1] algorithm comes from calculation of the needed Zak-transforms, and the $4L\left(1 + \frac{q}{p}\right)\log_2 N$ term comes from the transform to and from the Zak-domain. Compared to (22) and (23) this transformation uses longer FFTs. Algorithm 2 does away with the multiplication with complex exponentials in the [1] algorithm, and so the first term reduces to $L(8q)$.

Both the Poisson summation based algorithm and Algorithm 2 can do a DGT with $L \approx 2000000$ in less than 1 second on a standard PC at the time of writing. We have not created an efficient implementation of the algorithm from [1] in C so therefore we cannot reliably time it.

V. EXTENSIONS

The algorithm just developed can also be used to calculate the synthesis operator D_γ. This is done by applying Algorithm

Algorithm 3 Canonical Gabor dual window

GABDUAL(g, a, M)

1) $Phi =$WFAC(g, a, M)
2) **for** $r = \langle c \rangle$, $s = \langle d \rangle$
3) $\quad G \leftarrow Phi\,(:, :, r, s)$
4) $\quad Phi^d\,(:, :, r, s) \leftarrow \left(G \cdot G^T\right)^{-1} \cdot G$
5) **end for**
6) $g^d =$IWFAC$\left(Phi^d, a, M\right)$
7) **return** g^d

2 in the reverse order and inverting each line. The only lines that are not trivially invertible are lines 10-12, which becomes

10) $\Gamma \leftarrow Phi^d\,(:, :, r, s)$
11) $C \leftarrow Ctmp\,(:, :, s)$
12) $Psitmp\,(:, :, s) \leftarrow \Gamma \cdot C$

where the matrices $Phi^d\,(:, :, r, s)$ should be left inverses of the matrices $Phi\,(:, :, r, s)$ for each r and s.

The matrices $Phi^d\,(:, :, r, s)$ can be computed by Algorithm 1 applied to a dual Gabor window γ of the Gabor frame (g, a, M). It also holds that all dual Gabor windows γ of a Gabor frame (g, a, M) must satisfy that $Phi^d\,(:, :, r, s)$ are left inverses of the matrices $Phi\,(:, :, r, s)$. This criterion was reported in [11], [12].

A special left-inverse in the *Moore-Penrose pseudo-inverse*. Taking the pseudo-inverses of $Phi\,(:, :, r, s)$ yields the factorization associated with the canonical dual window of (g, a, M), [3]. This is Algorithm 3. Taking the polar decomposition of each matrix in $\Phi^g_{r,s}$ yields a factorization of the canonical tight window (g, a, M). For more information on these methods, as well as iterative methods for computing the canonical dual/tight windows, see [13].

VI. Special cases

We shall consider two special cases of the algorithm:

The first case is integer oversampling. When the redundancy is an integer then $p = 1$. Because of this we see that $c = a$ and $d = b$. This gives (16) the appearance

$$a = h_M q a - h_a a, \tag{26}$$

indicating that $h_M = 0$ and $h_a = -1$ solves the equation for all a and q. The algorithm simplifies accordingly, and reduces to the well known Zak-transform algorithm for this case, [10].

The second case is the short time Fourier transform. In this case $a = b = 1$, $M = N = L$, $c = d = 1$, $p = 1$, $q = L$ and as in the previous special case $h_M = 0$ and $h_a = -1$. In this case the algorithm reduces to the very simple and well known algorithm for computing the STFT.

VII. Implementation

The reason for defining the algorithm on multi-signals, is that the multiple signals can be handled at once in the matrix product in line 12 of Algorithm 2. This is a matrix product of two matrices size $q \times p$ and $p \times qW$, so the second matrix grows when multiple signals are involved. Doing it this way reuses the $\Phi^g_{r,s}$ matrices as much as possible, and this is an advantage on standard, general purpose computers with a deep memory hierarchy, see [5], [18].

The benefit of expressing Algorithm 2 in terms of loops (as opposed to using the Zak transform or matrix factorizations) is that they are easy to reorder. The presented Algorithm 2 is just one among many possible algorithms depending on in which order the r, s, k and l loops are executed. For a given platform, it is difficult a priory to estimate which ordering of the loops will turn out to be the fastest. The ordering of the loops presented in Algorithm 2 is the variant that uses the least amount of extra memory.

Implementations of the algorithms described in this paper can be found in the Linear Time Frequency Toolbox (LTFAT) available from http://ltfat.sourceforge.net. The implementations are done in both the Matlab/Octave scripting language and in C. A range of different variants of Algorithm 2 has been implemented and tested, and the one found to be the fastest on a small range of computers is included in the toolbox.

References

[1] M. J. Bastiaans and M. C. Geilen. On the discrete Gabor transform and the discrete Zak transform. 49(3):151–166, 1996.

[2] H. Bölcskei, F. Hlawatsch, and H. G. Feichtinger. Equivalence of DFT filter banks and Gabor expansions. In *SPIE 95, Wavelet Applications in Signal and Image Processing III*, volume 2569, part I, San Diego, july 1995.

[3] O. Christensen. Frames and pseudo-inverses. *J. Math. Anal. Appl.*, 195:401–414, 1995.

[4] O. Christensen. *An Introduction to Frames and Riesz Bases*. Birkhäuser, 2003.

[5] J. Dongarra, J. Du Croz, S. Hammarling, and I. Duff. A set of level 3 basic linear algebra subprograms. *ACM Trans. Math. Software*, 16(1):1–17, 1990.

[6] H. G. Feichtinger and T. Strohmer, editors. *Gabor Analysis and Algorithms*. Birkhäuser, Boston, 1998.

[7] H. G. Feichtinger and T. Strohmer, editors. *Advances in Gabor Analysis*. Birkhäuser, 2003.

[8] G. H. Golub and C. F. van Loan. *Matrix computations, third edition*. John Hopkins University Press, 1996.

[9] K. Gröchenig. *Foundations of Time-Frequency Analysis*. Birkhäuser, 2001.

[10] A. J. E. M. Janssen. The Zak transform: a signal transform for sampled time-continuous signals. *Philips Journal of Research*, 43(1):23–69, 1988.

[11] A. J. E. M. Janssen. On rationally oversampled Weyl-Heisenberg frames. pages 239–245, 1995.

[12] A. J. E. M. Janssen. The duality condition for Weyl-Heisenberg frames. In Feichtinger and Strohmer [6], chapter 1, pages 33–84.

[13] A. J. E. M. Janssen and P. L. Søndergaard. Iterative algorithms to approximate canonical Gabor windows: Computational aspects. *J. Fourier Anal. Appl.*, published online, 2007.

[14] S. Johnson and M. Frigo. A Modified Split-Radix FFT With Fewer Arithmetic Operations. *IEEE Trans. Signal Process.*, 55(1):111, 2007.

[15] P. Prinz. Calculating the dual Gabor window for general sampling sets. *IEEE Trans. Signal Process.*, 44(8):2078–2082, 1996.

[16] S. Qiu. Discrete Gabor transforms: The Gabor-gram matrix approach. *J. Fourier Anal. Appl.*, 4(1):1–17, 1998.

[17] T. Strohmer. Numerical algorithms for discrete Gabor expansions. In Feichtinger and Strohmer [6], chapter 8, pages 267–294.

[18] R. C. Whaley, A. Petitet, and J. Dongarra. Automated empirical optimization of software and the ATLAS project. Technical Report UT-CS-00-448, University of Tennessee, Knoxville, TN, Sept. 2000.

[19] Y. Y. Zeevi and M. Zibulski. Oversampling in the Gabor scheme. *IEEE Trans. Signal Process.*, 41(8):2679–2687, 1993.

Nonstationary Gabor Frames

Florent Jaillet [(1)], Peter Balazs [(1)] and Monika Dörfler [(1)]

(1) Acoustics Research Institute, Austrian Academy of Sciences, Wohllebengasse 12-14, A-1040 Vienna,Austria
florent@kfs.oeaw.ac.at, peter.balazs@oeaw.ac.at, monid@kfs.oeaw.ac.at

Abstract:

To overcome the limitation induced by the fixed time-frequency resolution over the whole time-frequency plane of Gabor frames, we propose a simple extension of the Gabor theory leading to the construction of frames with time-frequency resolution evolving over time or frequency. We describe the construction of such frames and give the explicit formulation of the canonical dual frame for some conditions. We illustrate the interest of the method on a simple example.

1. Introduction

Gabor analysis [7] is widely used for applications in signal processing. For some of these applications, which include processing of signals using Gabor frame multipliers [6, 1], the rigid construction of the Gabor atoms results in important limitations on the signal analysis and processing ability of the associated schemes. The Gabor transform uses time-frequency atoms built by translation over time and frequency of a unique prototype function, leading to a signal decomposition having a fixed time-frequency resolution over the whole time-frequency plane. This can be very restricting when dealing with signals with characteristics changing over the time-frequency plane. For example, this led some people to prefer the use of alternative decompositions with time-frequency resolution evolving with frequency in some applications, to better fit the feature of interest of the signal. Examples of such decompositions are the wavelet transform [5] or the decompositions using filter banks based on perceptive frequency scales for processing of audio signals, as for example gammatone filters [9].

A case for which the limitation induced by the constant time-frequency resolution of the Gabor transform can be seen is shown on the didactic example of Figure 1. On this figure, two spectrograms of the same glockenspiel signal are represented. These spectrograms are obtained by plotting the square absolute value of the Gabor coefficients using a color scale with a level coding in dB. Both spectrograms are obtained from the Gabor coefficients using the same type of window, but using two different window lengths. We see that the signal contains two very contrasting types of components:

- at the beginning of the notes, the signal presents sharp attacks which are spread in frequency, but very localized in time,
- during the resonance of the notes, the signal contains quasi-sinusoidal components which are spread in time, but very localized in frequency.

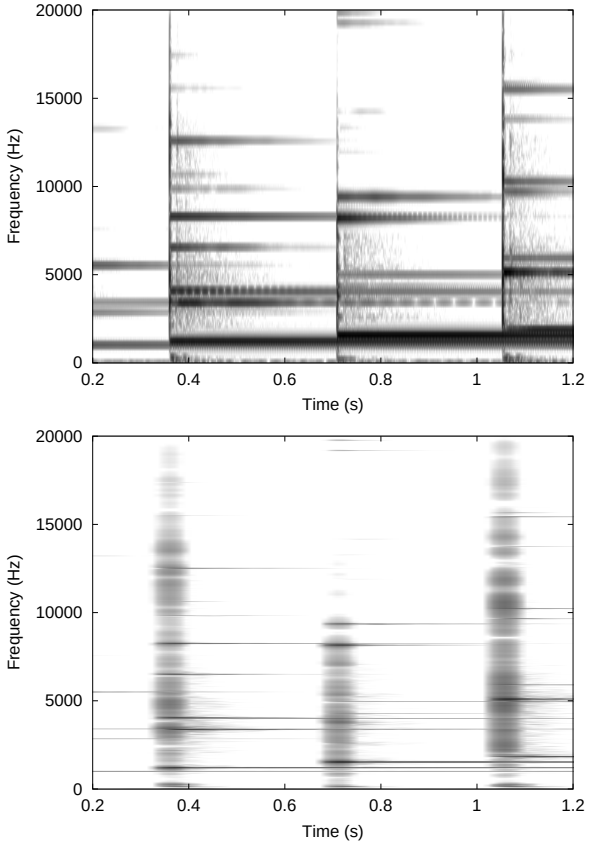


Figure 1: Two spectrograms of the same glockenspiel signal obtained using two different window lengths. On the top plot, a narrow window of 6 ms is used, on the bottom plot, a wide window of 93 ms is used.

We see that the use of the narrow window is well suited for the analysis and processing of the attacks, leading to a very sparse decomposition for these components, but gives an unsatisfying representation of the resonance, as the different sinusoidal components are not resolved. On the other hand, the wide window gives a good representation of the resonance part, but a blurred representation of the attacks. For this example, it appears that if we want to build an

optimised scheme for processing of both attacks and the resonances at the same time, it would be suitable to be able to adapt the time-frequency resolution locally for the different types of components.

The purpose of this paper is to describe one way to achieve this goal. For this, we show that, while staying in the context of frame theory [2, 4], the standard Gabor theory can be easily extended to provide some freedom of evolution of the time-frequency resolution of the decomposition in either time or frequency. Furthermore, this extension is well suited for applications as it can easily be implemented using fast algorithm based on fast Fourier transform [12].

We first describe the construction of the frames in Section 2., and then illustrate in Section 3. the potential of the approach on the preceding example of Figure 1.

2. Construction of the frames

2.1 Resolution evolving over time

As opposed to standard Gabor analysis, we replace time translation for the construction of atoms by the use of different windows for the different sampling positions in time. For each time position we still build atoms by regular frequency modulation. So using a set of functions $\{g_n\}_{n\in\mathbb{Z}}$ of $\mathbf{L}^2(\mathbb{R})$, for $m \in \mathbb{Z}$ and $n \in \mathbb{Z}$, we define atoms of the form:

$$g_{m,n}(t) = g_n(t)e^{i2\pi m b_n t}.$$

In practice we will choose each window g_n centered around a time a_n, and it will typically be constructed by translating a well localized window centered around 0 by a_n, as in the standard Gabor scheme, but with the possibility to vary the window g_n for each position a_n. Thus the sampling of the time-frequency plane is done on a grid which is irregular over time, but regular over frequency. Figure 2 shows an example of such a sampling grid. It can be noted that some results exist in Gabor theory for semi-regular sampling grids, as for example in [3]. Our study here uses a more general setting, as the sampling grid is in general not separable, and more importantly, the window can evolve over time.

In this case, the coefficients of the decomposition are given by:

$$c_{m,n} = \langle f, g_{m,n}\rangle,$$

and the frame operator is given by:

$$\mathbf{S}f = \sum_m \sum_n \langle f, g_{m,n}\rangle g_{m,n}.$$

The frame operator can be described by its kernel K given the following relation, which holds at least in a weak sense:

$$\mathbf{S}f(s) = \int K(t,s)f(t)\mathrm{d}t.$$

Here the kernel K simplifies according to the following relations:

$$\begin{aligned} K(t,s) &= \sum_m \sum_n \overline{g_n(t)} g_n(s) e^{i2\pi m b_n (s-t)} \\ &= \sum_n \overline{g_n(t)} g_n(s) \sum_m e^{i2\pi m b_n (s-t)} \\ &= \sum_n \frac{1}{b_n} \overline{g_n(t)} g_n(s) \sum_k \delta\left(s - t - \frac{k}{b_n}\right) \end{aligned}$$

thus,

$$\mathbf{S}f(s) = \sum_k \sum_n \frac{1}{b_n} \overline{g_n\left(s - \frac{k}{b_n}\right)} g_n(s)\, f\left(s - \frac{k}{b_n}\right)$$

In general, the inversion of $\mathbf{S}$ is not obvious. However we can identify a special case, which is analog to the "painless" case in standard Gabor analysis [8], for which the expression of $\mathbf{S}$ simplifies.

More precisely, we suppose from now on that for every $n \in \mathbb{Z}$, the function g_n has a limited time support $\operatorname{supp} g_n = [c_n, d_n]$ such that $d_n - c_n < \frac{1}{b_n}$. Due to this support condition, the terms of the summation over k in the preceding equation are 0 for $k \neq 0$ and the frame operator $\mathbf{S}$ becomes a multiplication operator:

$$\mathbf{S}f(s) = \sum_n \frac{1}{b_n} |g_n(s)|^2 f(s).$$

In this case the invertibility of the frame operator is easy to check and the system of functions $g_{m,n}$ forms a frame for $\mathbf{L}^2(\mathbb{R})$ if and only if $\forall t \in \mathbb{R}$, $\sum_n \frac{1}{b_n}|g_n(t)|^2 \simeq 1$.

When this condition is fulfilled, the canonical dual frame elements are given by:

$$\tilde{g}_{m,n}(t) = \frac{g_n(t)}{\sum_k \frac{1}{b_k}|g_k(t)|^2} e^{i2\pi m b_n t},$$

and the associated canonical tight frame elements can be calculated by:

$$\dot{g}_{m,n}(t) = \frac{g_n(t)}{\sqrt{\sum_k \frac{1}{b_k}|g_k(t)|^2}} e^{i2\pi m b_n t}.$$

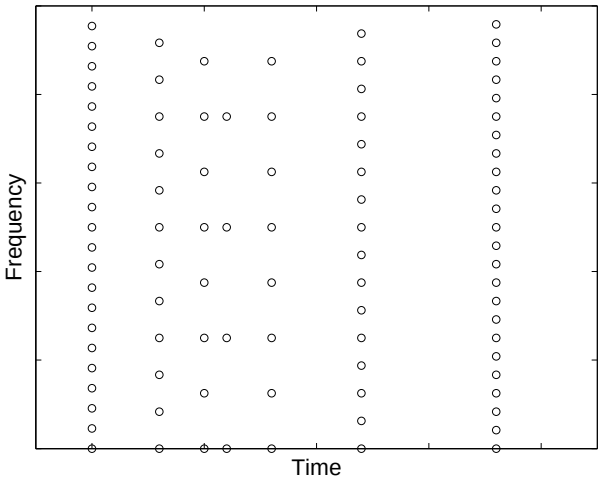


Figure 2: Example of sampling grid of the time-frequency plane when building a decomposition with time-frequency resolution evolving over time.

2.2 Resolution evolving over frequency

An analog construction is possible with a sampling of the time-frequency plane irregular over frequency, but regular over time. An example of the sampling grid in such a case is given on Figure 3.
In this case, we introduce a family of functions $\{h_m\}_{m\in\mathbb{Z}}$ of $\mathbf{L}^2(\mathbb{R})$, and for $m \in \mathbb{Z}$ and $n \in \mathbb{Z}$, we define atoms of the form:

$$h_{m,n}(t) = h_m(t - na_m).$$

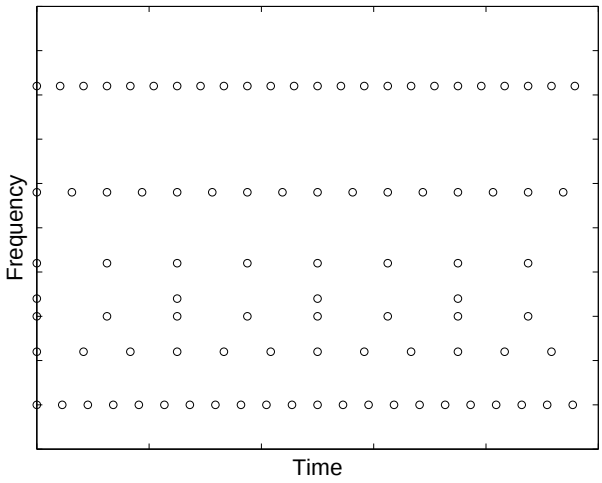


Figure 3: Example of sampling grid of the time-frequency plane when building a decomposition with time-frequency resolution evolving over frequency.

In practice we will choose each function h_m as a well localized pass-band function having a Fourier transform centered around some frequency b_n.
In this case the frame operator is given by:

$$\mathbf{T}f = \sum_m \sum_n \langle f, h_{m,n}\rangle h_{m,n},$$

and the problem is completely analog to the preceding up to a Fourier transform, as we have:

$$\widehat{\mathbf{T}f} = \sum_m \sum_n \langle \widehat{f}, \widehat{h_{m,n}}\rangle \widehat{h_{m,n}},$$

and $\widehat{h_{m,n}} = \widehat{h_m}(\nu)e^{-i2\pi n a_m \nu}$. So the preceding computation can be done, working on the Fourier transforms of the involved functions instead of directly on the functions.
Now the "painless" case appears when we suppose that for every $m \in \mathbb{Z}$, the function $\widehat{h_n}$ has a limited frequency support $\operatorname{supp} \widehat{h_n} = [e_n, f_n]$ such that $f_n - e_n < \frac{1}{a_n}$. Then the following expression holds:

$$\widehat{\mathbf{T}f}(\nu) = \sum_m \frac{1}{a_m} |\widehat{h_m}(\nu)|^2 \widehat{f}(\nu),$$

and the system of functions $h_{m,n}$ forms a frame of $\mathbf{L}^2(\mathbb{R})$ if and only if $\forall \nu \in \mathbb{R}$, $\sum_n \frac{1}{a_m} |\widehat{h_m}(\nu)|^2 \simeq 1$.
The associated canonical dual and tight frame can be computed as preceding, with the addition of an inverse Fourier transform.

2.3 Implementation

For the practical implementation, we have developed the equivalent theory in a finite discrete setting, that is to say working with complex vectors as signals. This theory won't be described here due to lack of space, but the construction is very similar to the one described in 2.1 and 2.2 and leads to a frame matrix which simplifies to a diagonal matrix in the "painless" case, suitable for applications.
The implementation is then very similar to the implementation of the standard Gabor case and can exploit fast Fourier transform algorithms for efficiency. The only differences compared to standard Gabor implementation are due to the fact that the storage of coefficients requires more advanced storage structures due to the irregularity of the time-frequency sampling grid, and that the computation of the dual window must be performed for every time position resulting in a slight increase in computational cost.

3. Example

The possibility to build a decomposition with time-frequency resolution evolving over time can be exploited to solve the problem described in example of Section 1. illustrated by Figure 1. For the corresponding glockenspiel signal, as we have seen before, the use of narrow window is suitable for the attacks of the notes, while a wide window should be used for the resonances. Figure 4 shows a representation built with our approach using a narrow window of 6 ms for the attacks and a wide window of 93 ms for the resonance. The frame used for this figure is a tight frame. It should be noticed that the evolution of the window size between the two target window lengths is smoothed in order to ensure that the atoms used for the decomposition maintain a "nice" shape, in the sense of having a good time-frequency concentration. This ensures the easy interpretability of the decomposition, especially for processing using frame multipliers.
This figure gives an idea of the type of decompositions that can be constructed with our approach and should be compared to the decomposition obtained using standard Gabor analysis on Figure 1. With our approach, it becomes possible to have a simultaneous good representation of both types of components of this signal while keeping the same processing ability than with standard Gabor.
We see that our approach allows to build decompositions with better time-frequency localization of the signal energy. This can be helpful for many processing tasks, in particular to reduce artifacts in component extraction or denoising.

4. Conclusion

Our approach enables the construction of frames with flexible evolution of time-frequency resolution over time or frequency. The resulting frames are well suited for applications as they can be implemented using fast algorithms, at a computational cost close to standard Gabor frames.
Exploiting evolution of resolution over time, the proposed approach can be of particular interest for applica-

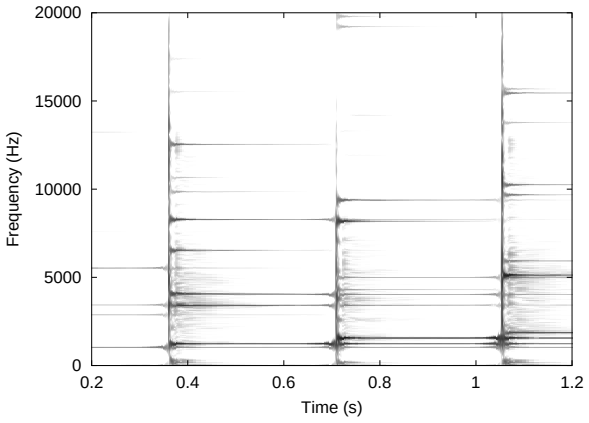


Figure 4: Spectrogram of the same glockenspiel signal as in Figure 1 using a nonstationary Gabor decomposition.

tions where the frequency characteristics of the signal are known to evolve significantly with time. Order analysis [11], in which the signal analyzed is produced by a rotating machine having evolving rotating speed, is an example of such application.

Exploiting evolution of resolution over frequency, the presented approach could be valuable for applications requiring the use of a tailored non uniform filter bank. In particular, it can be used to build filter banks following some perceptive frequency scale.

One difficulty when using our approach is to adapt the time-frequency resolution to the evolution of the signal characteristics. If prior knowledge is available, this can be done by hand, as for the example of Figure 4. But to go further, our approach could be extended to construct an adaptive decomposition of the signal by automatically adapting the resolution to the signal. To achieve this, we plan to investigate the possibility to couple our approach with the use of sparsity criterion as proposed in [10]. The general idea would then be to consider time segments of the signal, and for each time segment compare the sparsity criterion obtained for Gabor transforms computed with different possible windows. We would then use in our decomposition the window corresponding to the best criterion for each time segment, leading to a decomposition optimizing the sparsity of the decomposition over time.

Acknowledgment

This work was supported by the WWTF project MULAC ("Frame Multipliers: Theory and Application in Acoustics", MA07-025).

References:

[1] P. Balazs. Basic definition and properties of Bessel multipliers. *Journal of Mathematical Analysis and Applications*, 325(1):571585, January 2007.

[2] P. G. Casazza. The art of frame theory. *Taiwanese J. Math.*, 4(2):129–202, 2000.

[3] P. G. Casazza and O. Christensen. Gabor frames over irregular lattices. *Adv. Comput. Math.*, 18(2-4):329–344, 2003.

[4] O. Christensen. *An Introduction To Frames And Riesz Bases*. Birkhäuser, 2003.

[5] I. Daubechies. *Ten Lectures On Wavelets*. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM Philadelphia, 1992.

[6] H. G. Feichtinger and K. Nowak. A first survey of Gabor multipliers. In H. G. Feichtinger and T. Strohmer, editors, *Advances in Gabor analysis*, chapter 5, pages 99–128. Birkhäuser Boston, 2003.

[7] H. G. Feichtinger and T. Strohmer. *Gabor Analysis and Algorithms - Theory and Applications*. Birkhäuser Boston, 1998.

[8] K. Gröchenig. *Foundations of Time-Frequency Analysis*. Birkhäuser Boston, 2001.

[9] W. M. Hartmann. *Signals, Sounds, and Sensation*. Springer, 1998.

[10] F. Jaillet and B. Torrésani. Time-frequency jigsaw puzzle: adaptive multiwindow and multilayered gabor representations. *International Journal for Wavelets and Multiresolution Information Processing*, 5(2):293–316, 2007.

[11] H. Shao, W. Jin, and S. Qian. Order tracking by discrete Gabor expansion. *IEEE Transactions on Instrumentation and Measurement*, 52(3):754–761, 2003.

[12] J. S. Walker. *Fast Fourier Transforms*. CRC Press, 1991.

A Nonlinear Reconstruction Algorithm from Absolute Value of Frame Coefficients for Low Redundancy Frames

Radu Balan

Department of Mathematics, CSCAMM and ISR, University of Maryland, College Park, MD 20742, USA
rvbalan@math.umd.edu

Abstract:

In this paper we present a signal reconstruction algorithm from absolute value of frame coefficients that requires a relatively low redundancy. The basic idea is to use a nonlinear embedding of the input signal Hilbert space into a higher dimensional Hilbert space of sesquilinear functionals so that absolute values of frame coefficients are associated to relevant inner products in that space. In this space the reconstruction becomes linear and can be performed in a polynomial number of steps.

1. Introduction

Let us denote by $\mathbf{E}^n$ the n-dimensional space of signals (e.g. $E^n = \mathbf{R}^n$ or $E^n = \mathbf{C}^n$), and assume we are given a frame of m vectors $\{f_1, \ldots, f_m\} \subset \mathbf{E}^n$ that span E^n. Thus necessarily $m \geq n$. In this paper we look at the following problem: Given $c_l = |\langle x, f_l\rangle|$, $1 \leq l \leq m$, reconstruct the original signal $x \in \mathbf{E}^n$ up to a constant phase ambiguity, that is, obtain a signal $y \in \mathbf{E}^n$ such that $y = e^{i\varphi}x$ for some $\varphi \in [0, 2\pi)$.

This problem arises in several areas of signal processing (see [BCE06] for a more detailed discussion of these issues). In particular, in X-Ray Crystallography (see [LFB87]) it is known as the *phase retrieval problem*. In speech processing it is related to the use of cepstral coefficients in Automatic Speech Recognition as well as direct reconstruction from denoised spectogram (see [NQL82]). By the same token the solution posed here can be viewed as a new, nonlinear signal generating model.

Recently ([BBCE09]) we proposed a quasi-linear reconstruction algorithm that requires the frame to have high redundancy ($m = O(n^2)$). The algorithm works as follows. First note that two vectors $x, y \in \mathbf{E}^n$ that are equivalent (i.e. equal to one another up to a constant phase) generate the same rank-one operators $K_x = K_y$, where

$$K_u : \mathbf{E}^n \to \mathbf{E}^n \ , \ K_u(z) = \langle z, u\rangle u \tag{1}$$

with $u = x$ or $u = y$. Conversely, if $K_x = K_y$ then necessarily there exists a phase φ so that $y = e^{i\varphi}x$. Thus the reconstruction problem reduces to obtaining first K_x, and then a representative of the class $\hat{x}$. Next notice that the absolute value of frame coefficient $|\langle x, f_l\rangle|$ is related to the Hilbert-Schmidt inner product between K_x and K_{f_l}:

$$\langle K_x, K_{f_l}\rangle := trace(K_x K_{f_l}^*) = |\langle x, f_l\rangle|^2$$

Hence, if $\{K_{f_l} \ , \ 1 \leq l \leq m\}$ form a frame for the set of Hilbert-Schmidt operators (this is the same as the set of quadratic forms), then K_x can be reconstructed from d_l^2 with a linear algorithm, from where a vector $y \in \hat{x}$ can be obtained. Explicitely, the algorithm is as follows: First denote by $\{\widetilde{K_l} : \mathbf{E}^n \to \mathbf{E}^n \ , \ 1 \leq l \leq m\}$ the canonical dual frame of $\{K_{f_l} \ , \ 1 \leq l \leq m\}$.

1. Compute:

$$K_x = \sum_{l=1}^{m} c_l^2 \widetilde{K_l} \tag{2}$$

2. Assume $e \in \mathbf{E}^n$, $\|e\| = 1$ is so that $\|K_x e\| \neq 0$. Then:

$$y = \frac{1}{\sqrt{\langle K_x(e), e\rangle}} K_x(e) \tag{3}$$

is a vector in $\mathbf{E}^n$ equivalent to x.

While very appealing from a computational perspective, this algorithm requires the set $\{K_{f_l} \ , \ 1 \leq l \leq m\}$ to be complete (spanning) in the Hilbert space of $n \times n$ quadratic forms. In the real case ($\mathbf{E} = \mathbf{R}$) this latter Hilbert space is of dimension $n(n+1)/2$. In the complex case ($\mathbf{E} = \mathbf{C}$) the dimension becomes n^2. Thus the algorithm requires the original frame set $\{f_l \ , \ 1 \leq l \leq m\}$ to have $m = O(n^2)$ vectors. In practice this requirement may not be feasible. Furthermore, in [BCE06] we obtained that generically $m \geq 4n-2$ should suffice in the complex case, and $n \geq 2n-1$ should suffice in the real case. In this paper we present an algorithm that applies to a generic frame set of $m = 5.394n - 4.394$ vectors in the complex case, and $m = 2n-1$ in the real case. The main ingredient of this algorithm is the nonlinear embeding of $\mathbf{E}^n$ into a linear space $\Lambda_{d,d}$ of (d,d)-sesquilinear symmetric forms where the absolute value of frame coefficients provide the inner products with a frame set.

2. Nonlinear Embeddings

Let $\mathbf{E}^n$ be the signal n-dimensional Hilbert space. Let $\mathcal{F} = \{f_1, \ldots, f_m\}$ be a spanning set of m vectors in $\mathbf{E}^n$. Its *redundancy* is $r = m/n \geq 1$. Fix an integer $d \geq 1$ which is going to measure the embedding *depth*. Let $\Lambda_{d,d}(\mathbf{E}^n)$ denote the linear space of (d,d)-sesquilinear functionals, that is

$$\Lambda_{d,d}(\mathbf{E}^n) = \{ \ \alpha : \underbrace{\mathbf{E}^n \times \cdots \mathbf{E}^n}_{2d} \to \mathbf{C} \ \} \tag{4}$$

where $\alpha(y_1,\dots,y_d,z_1,\dots,z_d)$ is linear in $y_1,\dots,y_d$, and antilinear in $z_1,\dots,z_d$. Note $\Lambda_{d,d}(\mathbf{E}^n)$ is a vector space of dimension n^{2d}. Let $\{e_k\,,\,1\le k\le n\}$ be an orthonormal basis of $\mathbf{E}^n$. For each $2d$-tuple $(k_1,\dots,k_{2d})$ of integers from $1,\dots,n$ (repetitions are allowed) define

$$\begin{aligned}\delta_{k_1,\dots,k_{2d}}(y_1,\dots,y_d,z_1,\dots,z_d) &= \langle y_1,e_{k_1}\rangle\cdots\langle y_d,e_{k_d}\rangle\cdot\\ &\quad\langle e_{k_{d+1}},z_1\rangle\cdots\langle e_{k_{2d}},z_d\rangle \end{aligned}\tag{5}$$

Note $\Delta=\{\delta_{k_1,\dots,k_{2d}}\,;\,1\le k_l\le n, 1\le l\le 2d\}$ forms a basis in $\Lambda_{d,d}(\mathbf{E}^n)$. We define an inner product on $\Lambda_{d,d}(\mathbf{E}^n)$ so that this basis is orthonormal. Consider two sesquilinear functionals in $\Lambda_{d,d}(\mathbf{E}^n)$:

$$\alpha(y_1,\dots,y_d,z_1,\dots,z_d)=\langle y_1,a_1\rangle\cdots\langle y_d,a_d\rangle\langle b_1,z_1\rangle\cdots\langle b_d,z_d\rangle$$

$$\beta(y_1,\dots,y_d,z_1,\dots,z_d)=\langle y_1,g_1\rangle\cdots\langle y_d,g_d\rangle\langle h_1,z_1\rangle\cdots\langle h_d,z_d\rangle$$

Then their inner product is defined as

$$\langle\alpha,\beta\rangle:=\langle g_1,a_1\rangle\cdots\langle g_d,a_d\rangle\langle b_1,h_1\rangle\cdots\langle b_d,h_d\rangle\tag{6}$$

Extend this binary operation to an inner product on $\Lambda_{d,d}(\mathbf{E}^n)$. With this inner product Δ becomes an orthonormal basis for the Hilbert space $\Lambda_{d,d}(\mathbf{E}^n)$.

Now we are ready to define the nonlinear embedding of the input Hilbert space $\mathbf{E}^n$ in $\Lambda_{d,d}(\mathbf{E}^n)$. This is given by the map $\Phi:\mathbf{E}^n\to\Lambda_{d,d}(\mathbf{E}^n)$

$$\begin{aligned}\Phi(x)(y_1,\dots,y_d,z_1,\dots,z_d) &= \langle y_1,x\rangle\cdots\langle y_d,x\rangle\cdot\\ &\quad\cdot\langle x,z_1\rangle\cdots\langle x,z_d\rangle\end{aligned}\tag{7}$$

Let $E_d=span(\Phi(\Lambda_{d,d}(\mathbf{E}^n)))$ be the linear span of the embedding. Note in general $E_d\subsetneq\Lambda_{d,d}(\mathbf{E}^n)$ unless $d=1$. Let P denote the orthogonal projection onto E_d, $P:\Lambda_{d,d}(\mathbf{E}^n)\to E_d$.

Define now the following sesquilinear functionals associated to the frame set $\mathcal{F}$. Fix $1\le j_1,\dots,j_d\le m$.

$$\begin{aligned}\psi_{j_1,\dots,j_d}(y_1,\dots,y_d,z_1,\dots,z_d) &= \langle y_1,f_{j_1}\rangle\cdots\langle y_d,f_{j_d}\rangle\cdot\\ &\quad\cdot\langle f_{j_1},z_1\rangle\cdots\langle f_{j_d},z_d\rangle\end{aligned}\tag{8}$$

Note there are m^d distinct such functionals, however the number of distinct projections onto E_d is much smaller. Notice

$$\langle\Phi(x),\psi_{j_1,\dots,j_d}\rangle=|\langle x,f_{j_1}\rangle|^2\cdots|\langle x,f_{j_d}\rangle|^2\tag{9}$$

Thus if $(k_1,\dots,k_d)$ is a permutation of $(j_1,\dots,j_d)$ then $P\psi_{k_1,\dots,k_d}=P\psi_{j_1,\dots,j_d}$. For converse we need to assume first that frame vectors belong to distinct equivalence classes (that is, for any two $1\le l<j\le m$ and any $a\in[0,2\pi)$, $f_l\ne e^{ia}f_j$). Then we get that $P\psi_{k_1,\dots,k_d}=P\psi_{j_1,\dots,j_d}$ if and only if $(k_1,\dots,k_d)$ is a permutation of $(j_1,\dots,j_d)$. Thus we obtain that for frames with frame vectors in distinct equivalence classes the set

$$\Psi=\{\psi_{j_1,\dots,j_d}\,,\,1\le j_1\le j_2\le\cdots\le j_d\le m\}\tag{10}$$

is a maximal set of sesquilinear functionals of type (8) that have distinct projections through P.

For our algorithm to work we need to assume:

Assumption A. The set $P\Psi:=\{P\psi\,,\,\psi\in\Psi\}$ is spanning in E_d.

In section 4. we analyze the dimensionality constraint $|P\Psi|\ge dim(E_d)$, and in section 5. we present numerical results supporting Assumption A for a generic frame.

3. The Reconstruction Algorithm

Under Assumption A, let us denote by $\{\widetilde{\psi_{j_1,\dots,j_d}}\,,\,1\le j_1\le\cdots\le j_d\le m\}$ the canonical dual frame to $P\Psi$. This dual frame allows us to recover $\Phi(x)$. Recall $\{e_1,\dots,e_n\}$ is an orthonormal basis of $\mathbf{E}^n$. Notice the following relations:

$$\Phi(x)(e_k,\dots,e_k) = |\langle x,e_k\rangle|^{2d}\tag{11}$$

$$\sum_{k=1}^{n}(\Phi(x)(e_k,\dots,e_k))^{1/d} = \|x\|^2\tag{12}$$

$$\Phi(x)(\underbrace{e_j,\dots,e_j}_{2d-1},e_k) = |\langle x,e_j\rangle|^{2d-2}\langle e_j,x\rangle\langle x,e_k\rangle\tag{13}$$

From (11) and (13) we obtain:

$$\langle x,e_k\rangle=\frac{\langle x,e_j\rangle}{|\langle x,e_j\rangle|}\frac{\Phi(x)(e_j,\dots,e_j,e_k)}{(\Phi(x)(e_j,\dots,e_j,e_j))^{(2d-1)/2d}}\tag{14}$$

The Reconstruction Algorithm is as follows.

Reconstruction Algorithm

Input: Coefficients $c_1=|\langle x,f_1\rangle|$, ... $c_m=|\langle x,f_m\rangle|$.

Step 0. If $\sum_{k=1}^m c_k^2=0$ then $y=0$ and stop. Otherwise continue.

Step 1. Construct the following sesquilinear functional

$$\alpha=\sum_{1\le j_1\le\cdots\le j_d\le m}c_{j_1}^2\cdots c_{j_d}^2\widetilde{\psi_{j_1,\dots,j_d}}\tag{15}$$

Step 2. Find a $1\le j_0\le n$ so that $\alpha(e_{j_0},\cdots,e_{j_0})>0$. This is possible due to (12). Set

$$\nu=\sqrt[2d]{\alpha(e_{j_0},\dots,e_{j_0})}\tag{16}$$

Step 3. Set

$$y=\frac{1}{\nu^{2d-1}}\sum_{k=1}^{n}\alpha(\underbrace{e_{j_0},\dots,e_{j_0}}_{2d-1},e_k)e_k\tag{17}$$

Summarizing all results obtained so far we obtain:

Theorem 3..1 *For every $x\in\mathbf{E}^n$ there is $z\in\mathbf{C}$ so that $|z|=1$ and the output of the Reconstruction Algorithm satisfies $x=zy$. Specifically $z=\frac{\langle x,e_{j_0}\rangle}{|\langle x,e_{j_0}\rangle|}$, with j_0 obtained in Step 2.*

4. Redundancy Constraint

In this section we analyse the necessary condition $|\Psi|\ge dim(E_d)$.

4.1 The Cardinal of Set Ψ

The set Ψ given in (10) has the same cardinal as

$$\{(k_1,\dots,k_d)\,,\,1\le k_1\le\cdots\le k_d\le m\}\tag{18}$$

Let us denote this number by $M_{m,d}$. In order to compute it, consider the following cardinal equivalent set:

$$\{(n_1,\dots,n_m)\,,\,0\le n_1,\dots,n_m\le d, n_1+\cdots+n_m=d\}\tag{19}$$

The bijective correspondence between d-tuples of (18) and m-tuples of (19) is given by the following interpretation: n_l is the number of times l is presented in the d-tuple $(k_1, \ldots, k_d)$. Then, one can obtain the following recursion:

$$M_{m+1,d} = \sum_{r=0}^{d} M_{m,d}$$

where we set $M_{m,0} = 1$. Since $M_{1,d} = 1$, one obtains by induction that:

$$M_{m,d} = \left(\begin{array}{c} m+d-1 \\ m-1 \end{array} \right) = \frac{m(m+1)\cdots(m+d-1)}{d!} \tag{20}$$

4.2 The Dimension of E_d

Recall E_d is the linear span of vectorx $\Phi(x)$ in $\Lambda_{d,d}(\mathbf{E}^n)$. Recall also that Δ whose n^{2d} vectors are defined in (5) is an orthonormal basis in $\Lambda_{d,d}(\mathbf{E}^n)$. Let us denote by $N_{n,d}$ the dimension of E_d. We will describe an orthonormal basis in E_d. Fix $t_1, \ldots, t_n \in \mathbf{C}$ and expand:

$$\Phi(t_1 e_1 + \cdots t_n e_n) = \sum_{1 \le k_1, \ldots, k_{2d} \le n} t_{k_1} \cdots t_{k_d} \overline{t_{k_{d+1}}} \cdots \overline{t_{k_{2d}}} \cdot \cdot \delta_{k_1, \ldots, k_{2d}} \tag{21}$$

We shall group together terms containing same t_k terms. The real case will be treated separately from the complex case.

To simplify the exposition, we introduce notation common to both cases. Let us denote by $\underline{k} = (k_1, \ldots, k_r)$ an ordered r-tuple of integers each from 1 to n, where the length r is equal to $2d$ (in the real case), or d (in the complex case). Let us denote by $\mathcal{P}_r$ the set of r-permutations, and by $\mathcal{P}_{\underline{k}}$ the quotient set $\mathcal{P}_{\underline{k}} = \mathcal{P}/\sim_{\underline{k}}$ where $\pi', \pi'' \in \mathcal{P}_r$ are equivalent $\pi' \sim_{\underline{k}} \pi''$ if and only if $\pi'(\underline{k}) = \pi''(\underline{k})$. Note

$$|\mathcal{P}_{\underline{k}}| = \frac{r!}{m_1! \cdots m_n!}$$

where m_l denotes the number of repetitions of l in $\underline{k}$.

The Complex Case

In the complex case, t_k and $\overline{t_k}$ can be treated as indepedent (real) variables. Then terms in (21) are grouped using two independent d-tuples, $\underline{j} = (j_1, \ldots, j_d)$ and $\underline{l} = (l_1, \ldots, l_d)$ as follows

$$\sum_{1 \le j_1 \le \cdots \le j_d \le n} \sum_{1 \le l_1 \le \cdots \le l_d \le n} t_{j_1} \cdots t_{j_d} \overline{t_{l_1}} \cdots \overline{t_{l_d}} \times \times \sum_{\pi \in \mathcal{P}_{\underline{j}}} \sum_{\rho \in \mathcal{P}_{\underline{l}}} \delta_{\pi(j_1), \ldots, \pi(j_d), \rho(l_1), \ldots, \rho(l_d)}$$

Then the following sesquilinear functionals are orthonormal and form a basis in E_d:

$$d_{\underline{j},\underline{l}} = \frac{1}{\sqrt{|\mathcal{P}_{\underline{j}}|}\sqrt{|\mathcal{P}_{\underline{l}}|}} \sum_{\pi \in \mathcal{P}_{\underline{j}}} \sum_{\rho \in \mathcal{P}_{\underline{l}}} \delta_{\pi(j_1), \ldots, \pi(j_d), \rho(l_1), \ldots, \rho(l_d)} \tag{22}$$

Their number (and hence dimension of E_d) is equal to the number of ordered d-tuples $\underline{j}$ times the number of ordered d-tuples $\underline{l}$:

$$N_{n,d} = (M_{n,d})^2 = \left(\frac{n(n+1)\cdots(n+d-1)}{d!} \right)^2 \tag{23}$$

where we used (20). Note $N_{n,1} = n^2$ and we recover the complex case considered in [BBCE09].

The Real Case

In the real case, t_k and $\overline{t_k}$ are the same variables. Then the independent terms in (21) are indexed by $2d$-tuples $\underline{k} = (k_1, \ldots, k_{2d})$ as follows:

$$\sum_{1 \le k_1 \le k_{2d} \le n} t_{k_1} \cdots t_{k_{2d}} \sum_{\pi \in \mathcal{P}_{\underline{k}}} \delta_{\pi(k_1), \ldots, \pi(k_{2d})}$$

and an orthonormal basis of E_d is given by the following vectors indexed by ordered $2d$-tuples $\underline{k}$:

$$d_{\underline{k}} = \frac{1}{\sqrt{|\mathcal{P}_{\underline{k}}|}} \sum_{\pi \in \mathcal{P}_{\underline{k}}} \delta_{\pi(k_1), \ldots, \pi(k_{2d})} \tag{24}$$

The dimension of E_d in real case is then:

$$N_{n,d} = M_{n,2d} = \frac{n(n+1)\cdots(n+2d-1)}{(2d)!} \tag{25}$$

Note $N_{n,1} = \frac{n(n+1)}{2}$ and this recovers the real case in [BBCE09].

4.3 The Optimal Depth and Redundancy Condition

For given n we would like to find the minimum $m = m^*$ so that $M_{m,d} \ge N_{n,d}$ for some $d \ge 1$.

The Complex Case

We need to solve

$$\frac{m(m+1)\cdots(m+d-1)}{d!} \ge \left(\frac{n(n+1)\cdots(n+d-1)}{d!} \right)^2$$

or, completing the factorials:

$$(m+d-1)!\, d!\, ((n-1)!)^2 \ge (m-1)!\, ((n+d-1)!)^2$$

Let us denote

$$R(n,m,d) = \frac{(m+d-1)!d!((n-1)!)^2}{(m-1)!((n+d-1)!)^2} \tag{26}$$

Ideally we would like to solve:

$$\begin{array}{ll} (1) & d^*(n,m) = argmax_d R(n,m,d) \\ (2) & m^*(n) = min_{R(n,m,d^*(n,m)) \ge 1}\, m \end{array}$$

Instead we make the following choices for $d = d(n)$ and $m = m(n)$, and then optimize using Stirling's formula:

$$d = n-1 \tag{27}$$

$$m = A(n-1)+1. \tag{28}$$

Using Stirling's formula $n! = \sqrt{2\pi n} n^n e^{-n}$ we obtain for $R(n+1, An+1, n)$,

$$R(n+1, An+1, n) = \sqrt{\frac{8\pi(A+1)n}{A}} \left[\frac{A+1}{16} (1 + \frac{1}{A})^A \right]^n$$

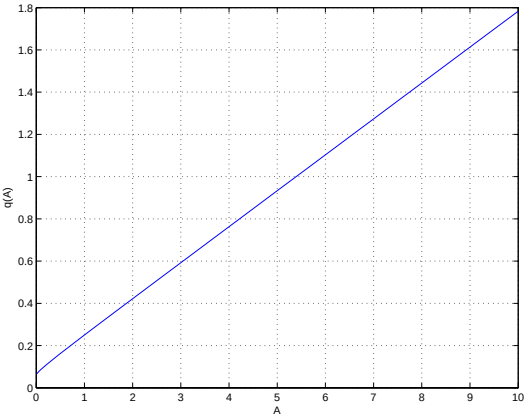


Figure 1: The plot of $q = q(A)$ from (29).

To obtain $R \geq 1$ for large n, we need

$$q(A) = \frac{A+1}{16}(1+\frac{1}{A})^A \geq 1 \tag{29}$$

In Figure 1 we plot the function $q = q(A)$. Numerically we obtain $A = 5.394$. The remaining factor in $R(n+1, An+1, n)$ becomes $5.376\sqrt{n} \geq 1$ for all n. Thus we obtain as sufficient conditions:

$$\begin{aligned} d &= n-1 && (30)\\ m &= 5.394n - 4.394 && (31)\end{aligned}$$

The Real Case
In the real case we need to solve

$$\frac{m(m+1)\cdots(m+d-1)}{d!} \geq \frac{n(n+1)\cdots(n+2d-1)}{(2d)!}$$

Following the same approach we obtain the following ratio function that we need to make supraunital:

$$R(n,m,d) = \frac{(m+d-1)!(n-1)!(2d)!}{(m-1)!(n+2d-1)!d!} \tag{32}$$

It follows:

$$R(n+1, 2n+1, n) = 1$$

Hence a possible choice is

$$\begin{aligned} d &= n-1 && (33)\\ m &= 2n-1 && (34)\end{aligned}$$

It is interesting to note that in the real case we recover the critical case $m \geq 2n-1$.

5. Numerical Evidence Supporting Genericity of the Assumption A.

While the previous section computed necessary conditions for Assumption A to hold true, we still need to prove (or check) that $P\Psi$ is frame in E_d. In this section we plot the distribution of eigenvalues of the frame operator associated to $P\Psi$ for one randomly generated example.
Using (22), each vector $P\psi_{\underline{k}}$ is represented by a $N_{n,d}$-vector whose components are indexed by a pair $(\underline{j},\underline{l})$, $F_{(\underline{j},\underline{l}),\underline{k}} = \langle \psi_{\underline{k}}, d_{\underline{j},\underline{l}} \rangle$. Explicitely this becomes

$$\begin{aligned} F_{(\underline{j},\underline{l}),\underline{k}} &= \frac{1}{\sqrt{|\mathcal{P}_{\underline{j}}|}\sqrt{|\mathcal{P}_{\underline{l}}|}} \sum_{\pi\in\mathcal{P}_{\underline{j}}} \sum_{\rho\in\mathcal{P}_l u} \langle e_{\pi(j_1)}, f_{k_1}\rangle \cdots \\ &\cdot\langle e_{\pi(j_d)}, f_{k_d}\rangle \langle f_{k_1}, e_{\rho(l_1)}\rangle \cdots \langle f_{k_d}, e_{\rho(l_d)}\rangle \end{aligned} \tag{35}$$

Thus $P\Psi$ is frame for E_d if and only if the $N_{n,d} \times M_{m,d}$ matrix F is of full rank. The frame operator is given by $S = FF^*$.
We considered the complex case ($\mathbf{E} = \mathbf{C}$) with the following parameters $n = 5$ and $d = 3$. For $m = 21$ the ratio function (26) takes the value $R(5, 21, 3) = 1.4457 > 1$. Note for the algorithm in [BBCE09] to work m has to be greater than or equal to n^2, that is $m \geq 25$. For a frame with 21 vectors in dimension 5 whose vectors are obtained as realizations of complex valued normal random variables of zero mean and variance 2 (each real and imaginary part is i.i.d. $\mathcal{N}(0,1)$), the distribution of eigenvalues of its frame operator is plotted in Figure 2. Note the conditioning number is $cond(S) = 6267.7$. While relatively large, the important thing to note is that the realization $P\Psi$ is frame (spanning) for E_d. While this result is by no

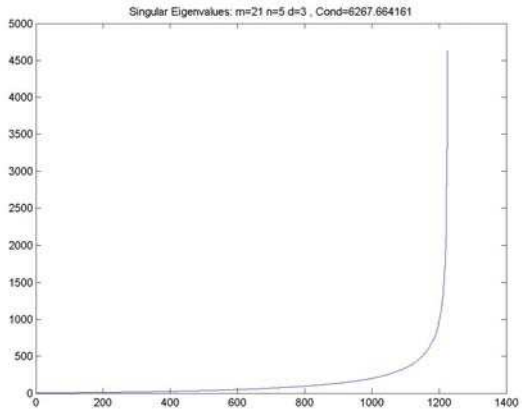


Figure 2: Distribution of eigenvalues for a random frame..

means a proof, or even an exhaustive experiment, it suggests the Assumption A might be generically true whenever $R(n,m,d) > 1$.

References:

[BCE06] R. Balan, P. Casazza, D. Edidin, *On signal reconstruction without phase*, Appl.Comput.Harmon.Anal. **20** (2006), 345–356.

[BBCE09] R. Balan, B. Bodman, P. Casazza, D. Edidin, *Painless reconstruction from magnitudes of frame coefficients*, to appear in the Journal of Fourier Analysis and Applications, 2009.

[LFB87] R. G. Lane, W. R. Freight, and R. H. T. Bates, *Direct Phase Retrieval*, IEEE Trans. ASSP **35**, no. 4 (1987), 520–526.

[NQL82] H. Nawab, T. F. Quatieri, and J. S. Lim, *Signal Reconstruction from the Short-Time Fourier Transform Magnitude*, in Proceedings of ICASSP 1984.

Matrix Representation of Bounded Linear Operators By Bessel Sequences, Frames and Riesz Sequence

Peter Balazs

Acoustics Research Institute, Austrian Academy of Sciences, Wohllebengasse 12-14, 1040 Wien, Austria.
peter.balazs@oeaw.ac.at

Abstract:

In this work we will investigate how to find a matrix representation of operators on a Hilbert space $\mathcal{H}$ with Bessel sequences, frames and Riesz bases as an extension of the known method of matrix representation by ONBs. We will give basic definitions of the functions connecting infinite matrices defining bounded operators on l^2 and operators on $\mathcal{H}$. We will show some structural results and give some examples. Furthermore in the case of Riesz bases we prove that those functions are isomorphisms. We are going to apply this idea to the connection of Hilbert-Schmidt operators and Frobenius matrices. Finally we will use this concept to show that every bounded operator is a generalized frame multiplier.

1. Introduction

From practical experience it became apparent that the concept of an orthonormal basis is not always useful. This led to the concept of frames, which was introduced by Duffin and Schaefer [12] and today it is one of the most important foundations of sampling theory [1].

The standard matrix description [8] of operators O using an ONB (e_k) is by constructing an matrix M with the entries $M_{j,k} = \langle Oe_k, e_j \rangle$. In [6] a concept was presented, where an operator R is described by the matrix $\left(\left\langle R\phi_j, \tilde{\phi}_i \right\rangle \right)_{i,j}$ with (ϕ_i) being a frame and $(\tilde{\phi}_i)$ its canonical dual. Such a kind of representation is used for the description of operators in [15] using Gabor frames and [19] using linear independent Gabor systems. In this work we are presenting the main ideas for Bessel sequences, frames and Riesz sequences and also look at the dual function which assigns an operator to a matrix. For proofs and details we refer to [3].

2. Notation and Preliminaries

2.1 Hilbert spaces and Operators

Let $\mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ denote the set of all linear and bounded operators from the Hilbert space $\mathcal{H}_1$ to $\mathcal{H}_2$. Furthermore we will denote the range of an operator A by $ran(O)$ and its kernel by $ker(A)$.

Let X, Y, Z be sets, $f : X \to Z$, $g : Y \to Z$ be arbitrary functions. The *Kronecker product* $\otimes_o : X \times Y \to Z$ is defined by $(f \otimes_o g)(x, y) = f(x) \cdot g(y)$. Let $f \in \mathcal{H}_1$, $g \in \mathcal{H}_2$ then define the *inner tensor product* as an operator from $\mathcal{H}_2$ to $\mathcal{H}_1$ by $(f \otimes_i \overline{g})(h) = \langle h, g \rangle f$ for $h \in \mathcal{H}_2$.

2.1.1 Hilbert Schmidt Operators

A bounded operator $T \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ is called a *Hilbert-Schmidt* ($\mathcal{HS}$) [18] operator if there exists an ONB $(e_n) \subseteq \mathcal{H}_1$ such that $\|T\|_{\mathcal{HS}} := \sqrt{\sum_{n=1}^{\infty} \|Te_n\|_{\mathcal{H}_2}^2} < \infty$. Let $\mathcal{HS}(\mathcal{H}_1, \mathcal{H}_2)$ denote the space of Hilbert Schmidt operators from $\mathcal{H}_1$ to $\mathcal{H}_2$.

2.2 Frames

A sequence $\Psi = (\psi_k | k \in K)$ is called a *frame* [5, 7] for the Hilbert space $\mathcal{H}$, if constants $A, B > 0$ exist, such that

$$A \cdot \|f\|_{\mathcal{H}}^2 \leq \sum_k |\langle f, \psi_k \rangle|^2 \leq B \cdot \|f\|_{\mathcal{H}}^2 \ \forall \ f \in \mathcal{H} \quad (1)$$

A sequence $\Psi = (\psi_k)$ is called a *Bessel sequence* with Bessel bound B if it fulfills the right inequality above. The index set will be omitted in the following, if no distinction is necessary.

A complete sequence (ψ_k) in $\mathcal{H}$ is called a *Riesz basis* if there exist constants $A, B > 0$ such that the inequalities

$$A \|c\|_2^2 \leq \left\| \sum_{k \in K} c_k \psi_k \right\|_{\mathcal{H}}^2 \leq B \|c\|_2^2$$

hold for all finite sequences (c_k).

3. Representing Operators with Frames

Let (ψ_k) be a frame in $\mathcal{H}_1$. An existing operator $U \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ is uniquely determined by its images of the frame elements. For $f = \sum_k c_k \psi_k$

$$U(f) = U(\sum_k c_k \psi_k) = \sum_k c_k U(\psi_k).$$

On the other hand, contrary to the case for ONBs, we cannot just choose a Bessel sequence (η_k) and define an operator just by choosing $V(\psi_k) := \eta_k$ and setting $V(\sum_k c_k \psi_k) = \sum_k c_k \eta_k$. This is in general not well-defined. Only if

$$\sum_k c_k \psi_k = \sum_k d_k \psi_k \Longrightarrow \sum_k c_k \eta_k = \sum_k d_k \eta_k$$

this definition is non-ambiguous, i.e. if $ker\,(D_{\psi_k}) \subseteq ker\,(D_{\eta_k})$. This condition is certainly fulfilled, if D_{ψ_k} is injective, i.e. for Riesz bases.
This problem can be avoided by using the following definition

$$V(f) := \sum_k \left\langle f, \tilde{\psi}_k \right\rangle \eta_k. \tag{2}$$

As (η_k) forms a Bessel sequence, the right hand side of Eq. (2) is well-defined. It is clearly linear, and it is bounded. The Bessel condition is necessary in the case of ONBs to get a bounded operator, too [8]. But contrary to the ONB case, here, in general, $V(\psi_k) \neq \eta_k$. So this option does not seem very useful. Instead of changing the sequence with which the coefficients are resynthesized, an operator can also be described by changing the coefficients, as presented in the following sections.

4. Matrix Representation

4.1 Motivation: Solving Operator Equalities

Given an operator equality $O \cdot f = g$ it is natural to discretize it to find a solution. Let $\Phi = (\phi_k)$ be a frame. Let us suppose that for a given g with coefficients $d = (d_k) = (\langle g, \phi_k \rangle)$ and a matrix representation M of O there is an algorithm to find the least square solution of

$$M \cdot c = d \tag{3}$$

for example using the pseudoinverse [7]. Still, if using frames, we can not expect to find a true solution for the operator equality just by applying $D_{\tilde{\Phi}}$ on c as in general c is not in $ran(C_\Phi)$ even if d is. But we see the following:

$$Of = g \Longleftrightarrow \sum_k \langle f, \phi_k \rangle O\tilde{\phi}_k = g \Longleftrightarrow$$

$$\Longleftrightarrow \sum_k \langle f, \phi_k \rangle \left\langle O\tilde{\phi}_k, \phi_k \right\rangle = \langle g, \phi_k \rangle$$

$$\Longleftrightarrow \mathcal{M}^{(\Phi,\tilde{\Phi})}(O) \cdot C_\Phi f = C_\Phi g.$$

It can be easily seen that this is equivalent to projecting c on $ran(C)$, solving $MC_\Phi D_{\tilde{\Phi}} c = d$, which is a common idea found in many algorithms, for example for a recent one see [20].

This gives us an algorithm for finding an approximative solution to the inverse operator problem $Of = g$.

1. Set $M = \mathcal{M}^{(\Phi,\tilde{\Phi})}(O)$.
2. Find a good finite dimensional approximation M_N of M by using the finite section method [14, 16] and
3. then apply an algorithm like e.g. the QR factorization [21] to find a solution for the operator equation.
4. and synthezise with the dual frame $\tilde{\Phi}$.

4.2 Bessel sequences

Theorem 4.2.1 *Let $\Psi = (\psi_k)$ be a Bessel sequence in $\mathcal{H}_1$ with bound B, $\Phi = (\phi_k)$ in $\mathcal{H}_2$ with B'.*

1. *Let $O : \mathcal{H}_1 \to \mathcal{H}_2$ be a bounded, linear operator. Then the infinite matrix*

$$\left(\mathcal{M}^{(\Phi,\Psi)}(O)\right)_{m,n} = \langle O\psi_n, \phi_m \rangle$$

defines a bounded operator from l^2 to l^2 with $\|\mathcal{M}\|_{l^2 \to l^2} \leq \sqrt{B \cdot B'} \cdot \|O\|_{\mathcal{H}_1 \to \mathcal{H}_2}$. As an operator $l^2 \to l^2$

$$\mathcal{M}^{(\Phi,\Psi)}(O) = C_\Phi \circ O \circ D_\Psi$$

This means the function $\mathcal{M}^{(\Phi,\Psi)} : \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2) \to \mathcal{B}(l^2, l^2)$ is a well-defined bounded operator.

2. *On the other hand let M be an infinite matrix defining a bounded operator from l^2 to l^2, $(Mc)_i = \sum_k M_{i,k} c_k$. Then the operator $\mathcal{O}^{(\Phi,\Psi)}$ defined by*

$$\left(\mathcal{O}^{(\Phi,\Psi)}(M)\right) h = \sum_k \left(\sum_j M_{k,j} \langle h, \psi_j \rangle \right) \phi_k,$$

$$\left\| \mathcal{O}^{(\Phi,\Psi)}(M) \right\|_{\mathcal{H}_1 \to \mathcal{H}_2} \leq \sqrt{B \cdot B'}\, \|M\|_{l^2 \to l^2}.$$

$$\mathcal{O}^{(\Phi,\Psi)}(M) = D_\Phi \circ M \circ C_\Psi = \sum_k \sum_j M_{k,j} \cdot \phi_k \otimes_i \overline{\psi}_j$$

This means the function $\mathcal{O}^{(\Phi,\Psi)} : \mathcal{B}(l^2, l^2) \to \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ is a well-defined bounded operator.

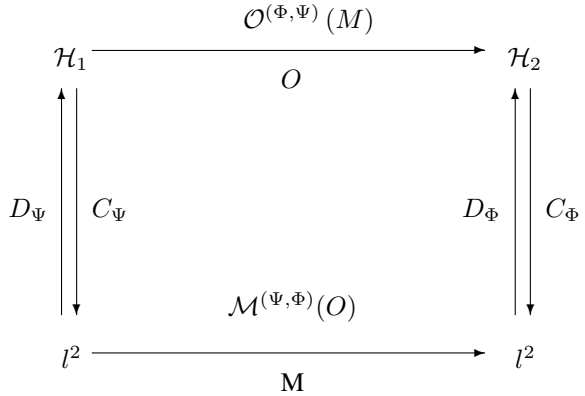


Figure 1: The operator induced by a matrix M and the matrix induced by an operator O.

If we do not want to stress the dependency on the frames and there is no change of confusion, the notation $\mathcal{M}(O)$ and $\mathcal{O}(M)$ will be used.
In the above theorem we have avoided the issue, when an infinite matrix defines a bounded operator from l^2 to l^2. A criterion has been proved in [9]:

Theorem 4.2.2 *An infinite matrix M defines a bounded operator from l^2 to l^2, if and only if $(M^*M)^n$ is defined for all $n = 1, 2, 3, \ldots$ and $\sup_n \sup_l \left| \left[(M^*M)^n_{l,l} \right]^{1/n} \right| < \infty$.*

For similar conditions see [17].

4.3 Frames

Proposition 4.3.1 *Let $\Psi = (\psi_k)$ be a frame in $\mathcal{H}_1$ with bounds A, B, $\Phi = (\phi_k)$ in $\mathcal{H}_2$ with A', B'. Then*

1. $\left(\mathcal{O}^{(\Phi,\Psi)} \circ M^{(\tilde{\Phi},\tilde{\Psi})}\right) = Id = \left(\mathcal{O}^{(\tilde{\Phi},\tilde{\Psi})} \circ M^{(\Phi,\Psi)}\right)$. *And therefore for all $O \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$:*
$$O = \sum_{k,j} \left\langle O\tilde{\psi}_j, \tilde{\phi}_k \right\rangle \phi_k \otimes_i \overline{\psi}_j$$
2. *$\mathcal{M}^{(\Phi,\Psi)}$ is injective and $\mathcal{O}^{(\Phi,\Psi)}$ is surjective.*
3. *Let $\mathcal{H}_1 = \mathcal{H}_2$, then $\mathcal{O}^{(\Psi,\tilde{\Psi})}(Id_{l^2}) = Id_{\mathcal{H}_1}$*
4. *Let $\Xi = (\xi_k)$ be any frame in $\mathcal{H}_3$, and $O : \mathcal{H}_3 \to \mathcal{H}_2$ and $P : \mathcal{H}_1 \to \mathcal{H}_3$. Then*
$$\mathcal{M}^{(\Phi,\Psi)}(O \circ P) = \left(\mathcal{M}^{(\Phi,\Xi)}(O) \cdot \mathcal{M}^{(\tilde{\Xi},\Psi)}(P)\right)$$

As a direct consequence we get the following corollary:

Corollary 4.3.2 *For the frame $\Phi = (\phi_k)$ the function $\mathcal{M}^{(\Phi,\tilde{\Phi})}$ is a Banach-algebra monomorphism between the algebra of bounded operators $(\mathcal{B}(\mathcal{H}_1, \mathcal{H}_1), \circ)$ and the infinite matrices of $\left(\mathcal{B}(l^2, l^2), \cdot\right)$.*

Lemma 4.3.3 *Let $O : \mathcal{H}_1 \to \mathcal{H}_2$ be a linear and bounded operator, let $\Psi = (\psi_k)$ and $\Phi = (\phi_k)$ be frames in $\mathcal{H}_1$ resp. $\mathcal{H}_2$. Then $\mathcal{M}^{(\Phi,\tilde{\Psi})}(O)$ maps $ran\,(C_\Psi)$ into $ran\,(C_\Phi)$ with*
$$\left(\langle f, \psi_k\rangle\right)_k \mapsto \left(\langle Of, \phi_k\rangle\right)_k .$$
If O is surjective, then $\mathcal{M}^{(\Phi,\tilde{\Psi})}(O)$ maps $ran\,(C_\Psi)$ onto $ran\,(C_\Phi)$. If O is injective, $\mathcal{M}^{(\Phi,\tilde{\Psi})}(O)$ is also injective.

The other function $\mathcal{O}$ is in general not so "well-behaved". It is, if the dual frames are biorthogonal. In this case these functions are isomorphisms, see the next section.

4.4 Riesz sequences

Theorem 4.4.1 *Let $\Phi = (\phi_k)$ be a Riesz basis for $\mathcal{H}_1$, $\Psi = (\psi_k)$ one for $\mathcal{H}_2$. The functions $\mathcal{M}^{(\Phi,\Psi)}$ and $\mathcal{O}^{(\tilde{\Phi},\tilde{\Psi})}$ between $\mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ and the infinite matrices in $\mathcal{B}(l^2, l^2)$ are bijective. $\mathcal{M}^{(\Phi,\Psi)}$ and $\mathcal{O}^{(\tilde{\Phi},\tilde{\Psi})}$ are inverse to each other. For $\mathcal{H}_1 = \mathcal{H}_2$ the identity is mapped on the identity by $\mathcal{M}^{(\Phi,\Psi)}$ and $\mathcal{O}^{(\tilde{\Phi},\tilde{\Psi})}$. If furthermore $\Psi = \Phi$ then $\mathcal{M}^{(\Phi,\tilde{\Phi})}$ and $\mathcal{O}^{(\Phi,\tilde{\Phi})}$ are Banach algebra isomorphisms, respecting the identities id_{l^2} and $id_{\mathcal{H}}$.*

5. Matrix Representation of $\mathcal{HS}$ Operators

We now have the adequate tools to state that $\mathcal{HS}$ operators correspond exactly to the Frobenius matrices, as expected. Let A be an m by n matrix, then $\|A\|_{fro} = \sqrt{\sum_{i=0}^{n-1} \sum_{j=0}^{m-1} |a_{i,j}|^2}$ is the *Frobenius norm*. Let us denote the set of all matrices with finite Frobenius norm by $l^{(2,2)}$, the set of *Frobenius matrices*.

Proposition 5.0.2 *Let $\Psi = (\psi_k)$ be a Bessel sequence in $\mathcal{H}_1$ with bound B, $\Phi = (\phi_k)$ in $\mathcal{H}_2$ with B'. Let M be a matrix in $l^{(2,2)}$. Then $\mathcal{O}^{(\Phi,\Psi)}(M) \in \mathcal{HS}(\mathcal{H}_1, \mathcal{H}_2)$, the Hilbert Schmidt class of operators from $\mathcal{H}_1$ to $\mathcal{H}_2$, with $\|\mathcal{O}(M)\|_{\mathcal{HS}} \leq \sqrt{BB'}\,\|M\|_{fro}$.*
Let $O \in \mathcal{HS}$, then $\mathcal{M}^{(\Phi,\Psi)}(O) \in l^{(2,2)}$ with $\|\mathcal{M}(O)\|_{fro} \leq \sqrt{BB'}\,\|O\|_{\mathcal{HS}}$.

5.1 Matrices and the Kernel Theorems

For $L^2(\mathbb{R}^d)$ the $\mathcal{HS}$ operators are exactly those integral operators with kernels in $L^2\left(\mathbb{R}^{2d}\right)$ [18]. This means that there exists a $\kappa_O \in L^2(\mathbb{R}^{2d})$ such an operator can be described as
$$(Of)(x) = \int \kappa_O(x, y) f(y) dy$$
Or in weak formulation
$$\langle Of, g\rangle = \int\int \kappa_O(x, y) f(y) \overline{g}(x) dy dx = \langle \kappa_O, f \otimes_o \overline{g}\rangle . \tag{4}$$
From 4.2.1 we know that
$$O = \sum_{j,k} \left\langle O\tilde{\psi}_j, \tilde{\phi}_k \right\rangle \phi_k \otimes_i \overline{\psi}_j$$
and so

Corollary 5.1.1 *Let $O \in \mathcal{HS}\left(L^2\left(\mathbb{R}^d\right)\right)$. Let $\Psi = (\psi_j)$ and $\Phi = (\phi_k)$ be frames in $L^2\left(\mathbb{R}^d\right)$. Then the kernel of O is given as:*
$$\kappa_O = \sum_{j,k} \mathcal{M}^{(\tilde{\Psi},\tilde{\Phi})}(O)_{k,j} \cdot \phi_k \otimes_o \overline{\psi}_j$$

This directly leads to the next concept.

6. Generalized Bessel Multipliers

Let m be a sequence and $\mathfrak{diag}(m)$ the matrix that has this sequence as diagonal. Then define
$$\mathbf{M}_{m,\Phi,\Psi} := \mathcal{O}^{(\Phi,\Psi)}(\mathfrak{diag}(m)) = \sum_k m_k \cdot \phi_k \otimes \psi_k$$
This means we have arrived quite naturally at the definition of frame multipliers as introduced in [2].
It is a very natural idea to extend this definition to include more side-diagonals:

Definition 6.0.2 *Let $\mathcal{H}_1$, $\mathcal{H}_2$ be Hilbert-spaces, let $(\psi_k)_{k\in L} \subseteq \mathcal{H}_1$ and $(\phi_k)_{k\in K} \subseteq \mathcal{H}_2$ be Bessel sequences. Let M be a $(K \times L)$-matrix that defines a bounded operator from l^2 to l^2. Define the operator $\mathbf{M}_{M,(\phi_k),(\psi_k)} : \mathcal{H}_1 \to \mathcal{H}_2$, the* generalized Bessel multiplier *for the Bessel sequences (ψ_k) and (ϕ_k), as the operator*

$$\mathbf{M}_{m,(\phi_k),(\psi_k)}(f) = \sum_l \sum_k M_{l,k} \langle f, \psi_k \rangle \phi_l.$$

The sequence m is called the symbol *of* $\mathbf{M}$. *If the sequence is a frame, we call the operator a 'generalized frame multiplier'.*

For Gabor frames, this is a particular case of the 'generalized Gabor multipliers' as found in [10] or [11] in this volume.
Using the results above we can write

Proposition 6.0.3 *For two frames $(\psi_k) \subseteq \mathcal{H}_1$ and $(\phi_k\} \subseteq \mathcal{H}_2$ every operator $O : \mathcal{H}_1 \to \mathcal{H}_2$ can be written as a generalized frame multiplier with the symbol $M_{l,k} = \left\langle O\tilde{\psi}_k, \tilde{\phi}_l \right\rangle$.*

Further results as the following are easy to prove:

Theorem 6.0.4 *Let $\mathbf{M} = \mathbf{M}_{m,\phi_k,\psi_k}$ be a Bessel multiplier for the Bessel sequences $(\psi_k) \subseteq \mathcal{H}_1$ and $(\phi_k\} \subseteq \mathcal{H}_2$ with the bounds B and B'. Then*

1. *If $M, M^* \in l^{1,\infty}$ with $\|M\|_{1,\infty} = K_1$ and $\|M^*\|_{1,\infty} = K_2$ then $\mathbf{M}$ is a well defined bounded operator with $\|\mathbf{M}\|_{Op} \leq \sqrt{B'BK_1K_2}$.*
2. *If $\sup_n \left\|M^{(n)}\right\|_{Op} = K < \infty$ then $\mathbf{M}$ is a well defined bounded operator with $\|\mathbf{M}\|_{Op} \leq \sqrt{B'B}K$.*
3. *If $(M^*M)^n$ is defined for $n = 1, 2, \dots$ and $\sup_n \sup_i \left[\langle M^*M)^n_{i,i}\right]^{1/n} = K < \infty$ then $\|\mathbf{M}\|_{Op} \leq \sqrt{B'B}K$.*
4. *If $\phi_k = \psi_k$ and $M \in \mathcal{B}(l^2)$ is a positive matrix, $\mathbf{M}$ is positive.*
5. *Let $M \in \mathcal{B}(l^2)$, then $\left(\mathbf{M}_{M,(\phi_k),(\psi_k)}\right)^* = \mathbf{M}_{M^*,(\psi_k),(\phi_k)}$. Therefore if M is self-adjoint and $\phi_k = \psi_k$, $\mathbf{M}$ is self-adjoint.*
6. *Let $M \in \mathcal{B}(l^2)$ be a matrix such that $\lim_n \left\|M^{(n)} - M\right\|_{Op} = 0$, then $\mathbf{M}$ is compact.*
7. *If $M \in l^{2,2}$, $\mathbf{M}$ is a Hilbert Schmidt operator with $\|M\|_{\mathcal{HS}} \leq \sqrt{B'}\sqrt{B}\ \|M\|_{2,2}$.*

Here for an operator A we denote $A^{(n)} = P_nAP_n$, where $P_n(x_0, x_1, x_2, \dots) = (x_1, x_2, \dots, x_{n-1}, 0, 0, \dots)$, see [14] (finite sections).

7. Perspectives

In this work we have investigated the basic idea of matrix representations using frames. An interesting question, as discussed in Section 4.1, is how to find a good finite approximation matrix. For first ideas in the Gabor case see [13, 10, 11, 22, 4].

8. Acknowledgments

The author would like to thank Jean-Pierre Antoine for many helpful comments and suggestions.
This work was partly supported by the WWTF project MULAC (Frame Multipliers: Theory and Application in Acoustics, MA07-025).

References:

[1] A. Aldroubi and K. Gröchenig. Non-uniform sampling and reconstruction in shift-invariant spaces. *SIAM Review*, 43:585–620, 2001.
[2] P. Balazs. Basic definition and properties of Bessel multipliers. *Journal of Mathematical Analysis and Applications*, 325(1):571–585, January 2007.
[3] P. Balazs. Matrix-representation of operators using frames. *Sampling Theory in Signal and Image Processing (STSIP)*, 7(1):39–54, Jan. 2008.
[4] J. Bendetto and G. Pfander. Frame expansions for Gabor multipliers. *Applied and Computational Harmonic Analysis (ACHA).*, 20(1):26–40, Jan. 2006.
[5] P. G. Casazza. The art of frame theory. *Taiwanese J. Math.*, 4(2):129–202, 2000.
[6] O. Christensen. Frames and pseudo-inverses. *J. Math. Anal. Appl*, 195(2):401–414, 1995.
[7] O. Christensen. *An Introduction To Frames And Riesz Bases*. Birkhäuser, 2003.
[8] J. B. Conway. *A Course in Functional Analysis*. Graduate Texts in Mathematics. Springer New York, 2. edition, 1990.
[9] Lawrence Crone. A characterization of matrix operator on l^2. *Math. Z.*, 123:315–317, 1971.
[10] M. Dörfler and B. B. Torrésani. Spreading function representation of operators and gabor multiplier approximation. In *Proceedings of SAMPTA'07*, 2007.
[11] M. Dörfler and B. B. Torrésani. Representation of operators by sampling in the time frequency domain. In *Proceedings of SAMPTA'09*, 2009.
[12] R. J. Duffin and A. C. Schaeffer. A class of nonharmonic Fourier series. *Trans. Amer. Math. Soc.*, 72:341–366, 1952.
[13] H. G. Feichtinger, M. Hampejs, and G. Kracher. Approximation of matrices by Gabor multipliers. *IEEE Signal Procesing Letters*, 11(11):883–886, 2004.
[14] I. Gohberg, S. Goldberg, and M. Kaashoek. *Basic Classes of Linear Operators*. Birkhäuser, 2003.
[15] K. Gröchenig. Time-frequency analysis of Sjöstrand's class. *Rev. Mat. Iberoam.*, 22:(to appear), 2006.
[16] O.Christensen and T.Strohmer. The finite section method and problems in frame theory. *Journal of Approximation Theory*, 133(2):221–237, 2005.
[17] W. H. Ruckle. *Sequence spaces*. Research Notes in Mathematics 49. Pitman London, 1981.
[18] R. Schatten. *Norm Ideals of Completely Continious Operators*. Springer Berlin, 1960.
[19] T. Strohmer. Pseudodifferential operators and Banach algebras in mobile communications. *Appl.Comp.Harm.Anal.*, 20(2):237–249, 2006.
[20] G. Teschke. Multi-frame representations in linear inverse problems with mixed multi-constraints. *Applied and Computational Harmonic Analysis*, 22(1):43–60, Jan. 2007. DFG-SPP-1114 preprint 90.
[21] L. N. Trefethen and D. Bau III. *Numerical Linear Algebra*. SIAM Philadelphia, 1997.
[22] P. Wahlberg and P Schreier. Gabor discretization of the Weyl product for modulation spaces and filtering of non-stationary stochastic processes. *Appl.Comp.Harm.Anal.*, 26:97–120, 2009.

Quasi-Random Sequences for Signal Sampling and Recovery

Mirosław Pawlak [1] and Ewaryst Rafajłowicz [2]

(1) Dept. of Electrical & Computer Eng., University of Manitoba, Winnipeg, Manitoba, Canada, R3T 2N2
(2) Institute of Computer Eng., Control and Robotics, Wrocław University of Technology, Wroclaw, Poland
pawlak@ee.umanitoba.ca, ewaryst.rafajlowicz@pwr.wroc.pl

Abstract:

The problem of reconstruction of band-limited signals from sampled and noisy observations is studied. It is proposed to sample a signal at quasi-random points, that form a deterministic sequence with properties resembling a random variable being uniformly distributed. Such quasi-random points can be easily and efficiently generated yielding signal reconstruction algorithms with the improved accuracy. In fact, in this paper we propose a reconstruction method based on the modified orthogonal sampling formula where the sampling rate and the reconstruction rate are treated separately. This distinction is necessary to ensure consistency of the reconstruction algorithm in the presence of noise. Asymptotical properties of the algorithm are evaluated including its convergence to the true signal and the corresponding rate. It is shown that the rate of convergence is better than that for reconstructions algorithms that utilize the traditional uniform sampling. Similar results are also obtained for the case of multivariate signals.

1. Introduction

Signal sampling is an inherent part of the modern signal processing theory and as such it has attracted a great deal of research activities lately [9], [10]. In particular, the problem of signal sampling and recovery from imperfect data has been addressed in a number of recent works [1], [2], [7]. In this case, one assumes that the signal samples $\{f(k\tau)\}$ are observed with noise, i.e., we have

$$y_k = f(k\tau) + z_k,$$

where z_k is uncorrelated noise process with $E\, z_k = 0$, $\mathrm{var}(z_k) = \sigma^2 < \infty$. Throughout the paper we assume that $f(t)$ has a bounded spectrum and that $f(t)$ is a finite energy type signal. Any signal with such a property is referred to as band-limited and will denote this class of signals as $BL(\Omega)$, where Ω is the bandwidth of $f(t)$.
The celebrated Whittaker-Shannon theorem says that any band-limited signal $f(t)$ can be perfectly recovered from its discrete values $\{f(k\tau)\}$ provided that $\tau \leq \pi/\Omega$. Application of the resulting interpolation formula to noisy data would lead to the following reconstruction scheme based on $2n+1$ random samples

$$f_n(t) = \sum_{|k|\leq n} y_k \,\mathrm{sinc}\left(\pi\tau^{-1}\,(t - k\tau)\right), \qquad (1)$$

where $\mathrm{sinc}(t) = \sin(t)/t$, and $\tau \leq \pi/\Omega$. The fundamental question, which arises is whether $f_n(t)$ can be a consistent estimate of $f(t)$ for any $f \in BL(\Omega)$. Hence, whether $\varrho(f_n, f) \to 0$ as $n \to \infty$, in a certain probabilistic sense, for some distance measure ϱ. Since $f(t)$ is assumed to be square integrable, then the natural measure between $f_n(t)$ and $f(t)$ is the mean integrated square error

$$MISE(f_n) = \mathbf{E} \int_{-\infty}^{\infty} (f_n(t) - f(t))^2\, dt. \qquad (2)$$

It can be easily shown, see [6], that $MISE(f_n) \to \infty$ as $n \to \infty$ for any fixed $\tau \leq \pi/\Omega$. This unpleasant property of the estimate $f_n(t)$ is caused by the presence of the noise process in the observed data and the fact that $f_n(k\tau) = y_k$, i.e., $f_n(t)$ interpolates the noisy observations. It is clear that one should avoid interpolation schemes in the presence of noise since they would retain random errors. The aim of this paper is to propose a consistent estimate of $f(t)$ being a smooth correction of the naive algorithm $f_n(t)$. This task is carried out by sampling a signal at irregularly spaced quasi-random points and by carefully selecting the number of terms in the sampling series. The conditions for consistency of our estimate are established and the corresponding rate of convergence is evaluated.
The statistical aspects of signal sampling and recovery have been examined first in [5], and next in [6], [7], [2], [1]. In [2], [1] the sampling rate τ has been assume to be a fixed constant. This assumption, however, cannot lead to consistent estimates of the true signal of the band-limited type. On the other hand, in [5], [6], [7] $\tau = \tau_n$

such that $\tau_n \to 0$ as $n \to \infty$ with a controlled rate. Such a choice of τ allows us to design a signal recovery algorithm for which the reconstruction error $MISE$ tends to zero with a certain rate. In this paper, we propose a nonlinear sampling scheme based on the theory of quasi-random sequences, i.e., we observe the following noisy samples

$$y_k = f(\tau_k) + z_k,$$

where $\{\tau_k\}$ is a sequence of quasi-random points. We show that a proper choice of $\{\tau_k\}$ leads to the reconstruction algorithm with the improved convergence rate.

2. Reconstruction Algorithms with Quasi-Random Points

The notion of quasi-random sequences has been originally established in the theory of numerical integration [4]. A sequence of real numbers $\{x_j\}$ is said to be a quasi-random sequence in $[0,\ 1]$ if for every continuous function $b(x)$ on $[0,\ 1]$ we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} b(x_j) = \int_0^1 b(x)\, dx. \tag{3}$$

Quasi-random sequences are also called equidistributed sequences, since (3) means that the sequence $\{x_j\}$ behaves like uniformly distributed random variables. Nevertheless, an important property of quasi-random sequences is that they are more uniform than random uniform sequences which tend to clump. A consequence of this fact is that the accuracy of approximating integrals based on quasi-random sequences is superior to the accuracy obtained by random sequences. In fact, the celebrated Koksma-Hlawka inequality [4] says that for any function of bounded variation on $[0, 1]$ we have

$$\left| n^{-1} \sum_{j=1}^{n} f(x_j) - \int_0^1 f(t)dt \right| \leq \mathcal{V}(f) D_n^*,$$

where $\mathcal{V}(f)$ is the total variation of f on $[0, 1]$, and D_n^* denotes the so-called discrepancy of the quasi-random sequence $\{x_j\}$. The discrepancy measures the strength of the sequence to approximate the uniform distribution on $[0, 1]$. There are quasi-random sequences with discrepancy of order $O(log(n)/n)$ [4]. This should be contrasted with a random sequence of uniformly distributed points on $[0, 1]$ that possesses the discrepancy of order $O(1/\sqrt{n})$. This basic observation plays a key role in our developments concerning the signal recovery problem from quasi-random points. Numerous quasi-random sequences have been constructed that have the aforementioned property of approximating the uniform distribution. The simplest, and sufficient for our purposes, way of generating a quasi-random sequence is the following

$$x_j = \text{frac}(j\,\vartheta), \tag{4}$$

where ϑ is an irrational number and frac(.) denotes the fractional part of a number in the parenthesis. A good choice of ϑ is $(\sqrt{5}-1)/2$, see [8] for an extensive discussion on the choice of ϑ.

Since band-limited signals are defined on the whole real line we need a rescaled version of quasi-random sequences. Thus, let us define the following sampling points on the interval $[-T,\ T]$

$$\tau_j = T\,\text{sgn}(j)\,\text{frac}(|j|\,\vartheta), \quad j = 0, \pm1, \pm2, \ldots,\ n, \tag{5}$$

where sgn(.) is the sign of a number. The observation horizon T must increase with n such that $T(n) \to \infty$ as $n \to \infty$. In order, however, to establish the consistency result of our reconstruction algorithm we must control the growth of $T(n)$. The approximation property of quasi-random sequences applied to the sequence defined in (5) reads now as

$$\frac{2T}{2n+1} \sum_{|j|\leq n} f(\tau_j) \approx \int_{-T}^{T} f(t)\, dt. \tag{6}$$

It has been known since the work of Hardy [3] that the cardinal expansion can be viewed as the orthogonal expansion in $BL(\Omega)$. Using this fact and then the reasoning as in [5] we can define the following estimate of $f(t)$

$$\tilde{f}_n(t) \;=\; \sum_{|k|\leq N} \tilde{c}_k s_k(t)\ , \tag{7}$$

$$\tilde{c}_k \;=\; \frac{2T}{(2n+1)\,h} \sum_{|j|\leq n} y_j\, s_k(\tau_j)\ , \tag{8}$$

where $\{\tau_j\}$ is the quasi-random sequence defined in (5). Here $\{s_k(t) = \text{sinc}(\pi h^{-1}(t-kh)), k = 0, \pm1, \ldots\}$ forms the orthogonal and complete system in $BL(\Omega)$ provided that $h \leq \pi/\Omega$. The corresponding Fourier coefficient is $c_k = h^{-1}\int_{-\infty}^{\infty} f(t)s_k(t)dt$. It is also clear that for $f \in BL(\Omega)$ we have $c_k = f(kh)$. The parameter h is called the reconstruction rate. In (7) the parameter N defines the number of terms in the expansion which are taken into account and $2n+1$ is the sample size. The truncation parameter plays important role in our asymptotic analysis, i.e., N depends on n such that $N(n) \to \infty$ with the controlled rate. It is also worth noting that the sampling rate is nonuniform (defined by the discrepancy of the quasi-random sequence in (5)) and different than the reconstruction rate h. We assume that h is constant and not greater than π/Ω.

Throughout the paper we use the worst localized base system utilizing the $sinc$ function. The methodology

presented in this paper can be extended to the windowed version of the estimate $\tilde{f}_n(t)$ of the form

$$\hat{f}_n(t) = \sum_{|k|\leq n} w_k \tilde{c}_k s_k(t),$$

where $\{w_k, |k| \leq n\}$ is a sequence of numbers such that $0 \leq w_k \leq 1$. The proper choice of this window sequence yields an estimate with better time-localized properties and consequently better convergence rates. The case when $w_k = 1$ for $|k| \leq N$ and $w_k = 0$ otherwise corresponds to the estimate $\tilde{f}_n(t)$.

3. The MISE Consistency and Rate

In this section we summarize the results concerning the convergence of MISE($\tilde{f}_n$) to zero as $n \to \infty$ for any signal $f \in BL(\Omega)$. Also the rate of convergence is established.
Due to Pareseval's formula we can decompose the $MISE(\tilde{f}_n)$ as follows:

$$MISE(\tilde{f}_n) = h \sum_{|k|\leq N} var(\tilde{c}_k) + h \sum_{|k|\leq N} (E\tilde{c}_k - c_k)^2 + h \sum_{|k|\geq N} c_k^2.$$

The first term of the decomposition controls the stochastic part of the estimate, whereas the the remaining term describe the systematic error (bias) of the estimate. A careful examination of these terms lead to the following result on the consistency of our estimate.

Theorem 1 *Let $f \in BL(\Omega)$ and let the reconstruction rate h be constant such that $h \leq \pi/\Omega$. Suppose that $N(n) < T(n)/h$. Assume $T(n) \to \infty$, $N(n) \to \infty$ such that $T(n)$ does not grow faster than $\sqrt{n}/log(n)$. Let, moreover,*

$$\frac{N(n)T(n)}{n} \to 0.$$

Then

$$MISE(\tilde{f}_n) \to 0$$

as $n \to \infty$.

The conditions required on the parameters $T(n)$ and $N(n)$ in Theorem 1 impose some general restrictions on their growth. In order further see how to choose $T(n)$ and $N(n)$ let us assume the following condition on the decay of band-limited signals.

(F) There exists $r \geq 0$ and a constant $C_f > 0$ such that for $|t|$ sufficiently large we have $|f(t)| \leq C_f/|t|^{r+1}$.

This assumption can be also expressed in the frequency domain by requiring that the Fourier transform of $f(t)$ has r derivatives on $[-\Omega, \Omega]$. A further analysis of the reconstruction error leads to the following bound

$$MISE(\tilde{f}_n) \leq (2\,N+1)\left(C_1 T^{-(2r+1)} + \frac{C_2 T^3 \log^2(n)}{n^2} + \frac{C_3 T}{n}\right) + C_4 N^{-(2r+1)}, \quad (9)$$

for some constants C_1, C_2, C_3, C_4. By optimizing the above bound we can obtain the following asymptotically optimal choice of $T(n)$ and $N(n)$.

$$T^*(n) = an^{\frac{1}{2r+3}} \quad N^*(n) = bn^{\frac{1}{2r+3}},$$

subject to the condition $a > bh$. Plugging these values of $T(n)$ and $N(n)$ back into the bound for $MISE(\tilde{f}_n)$ we obtain the following rate

$$MISE(\tilde{f}_n) = O(n^{-\frac{2r+1}{2r+3}}).$$

It is worth noting that under Assumption (F) the best possible rate obtained for the reconstruction algorithms discussed in [6] and [7] is of order $O(n^{-\frac{r}{r+1}})$. This is clearly a slower rate than the one obtained in this paper.

4. Concluding Remarks

In this paper we have proposed an algorithm for recovering a band-limited signal observed under noise. Assuming that the signal is a square integrable function the sufficient conditions for the convergence of the mean integrated square error have been established. The distinguishing feature of the proposed approach is its utilization of nonuniform samples taken at quasi-random points. When quasi-random sequences are applied to the problem of numerical evaluation of integrals they reveal the approximation rate $O(log(n)/n)$ for a class of bounded variation functions. This rate is superior to the rate $O(1/\sqrt{n})$ that characterizes usual numerical algorithms and classical Monte Carlo methods. This advantage of quasi-random sequences seems to be carried out to the problem of signal sampling and recovery. In our consistency results we assume that the reconstruction rate h is constant and could be chosen as large as π/Ω. One could also consider the case when $h = h(n)$ and $h(n) \to 0$ as $n \to \infty$. The estimates with variable h would be needed for the problem of recovering not necessarily band-limited signals. Finally, let us mention that the results of this paper can be extended to the d-dimensional case, where the orthogonal system can be obtained in the form of the product of sinc functions, i.e., $\mathbf{s_k}(\mathbf{t}) = \prod_{i=1}^{d} s_{k_i}(t_i)$, where $\mathbf{k} = (k_1, k_2, \ldots, k_d)$,

$\mathbf{t} = (t_1, t_2, \ldots, t_d)$. We should mention that multidimensional quasi-random sequences can be generated in a relatively straightforward way. Moreover, they exhibit the favorite discrepancy of order $O(n^{-1}(log(n))^d)$ for any d. This fact may have important consequences for sampling problems of two-dimensional objects like images.

5. Acknowledgements

The work of E. Rafajłowicz was supported by the Research and Development Grant from the Ministry of Science and Higher Education of Poland.

References:

[1] A. Aldroubi, C. Leonetti, and Q. Sun. Error analysis of frame reconstruction from noisy samples. *IEEE Trans. Signal Processing*, 56:2311–2315, 2008.

[2] Y.C. Eldar and M. Unser. Non-ideal sampling and interpolation from noisy observations in shift-invariant spaces. *IEEE Trans. Signal Processing*, 54:2636–2651, 2006.

[3] G.H. Hardy. Notes on special systems of orthogonal functions (iv): the orthogonal functions of Whittaker's cardinal series. *Proc. Camb. Phil. Soc.*, 37:331–348, 1941.

[4] L. Kuipers and H. Niederreiter. *Uniform Distribution of Sequences*. Wiley, New York, 1974.

[5] M. Pawlak and E. Rafajłowicz. On restoration of band-limited signals. *IEEE Trans. Information Theory*, 40:1490–1503, 1994.

[6] M. Pawlak, E. Rafajłowicz, and A. Krzyżak. Postfiltering versus prefiltering for signal recovery from noisy samples. *IEEE Trans. Information Theory*, 49:3195–3212, 2003.

[7] M. Pawlak and U. Stadtmüller. Signal sampling and recovery under dependent noise. *IEEE Trans. Information Theory*, 53:2526–2541, 2007.

[8] E. Rafajłowicz and R. Schwabe. Equidistributed designes in nonparametric regression. *Statistica Sinica*, 13:129–142, 2003.

[9] M. Unser. Sampling – 50 years after Shannon. *Proceedings of the IEEE*, 88:569–587, 2000.

[10] P.P. Vaidyanathan. Generalizations of the sampling theorems: seven decades after Nyquist. *IEEE Trans. on Ciruits and Systems – I : Fundamental Theory and Applications*, 48:1094–1109, 2001.

On the incoherence of noiselet and Haar bases

Tomas Tuma, Paul Hurley
IBM Research, Zurich Laboratory 8803 Rüschlikon, Switzerland
E-mail: {uma,pah}@zurich.ibm.com

Abstract:

Noiselets are a family of functions completely uncompressible using Haar wavelet analysis. The resultant perfect incoherence to the Haar transform, coupled with the existence of a fast transform has resulted in their interest and use as a sampling basis in compressive sampling. We derive a recursive construction of noiselet matrices and give a short matrix-based proof of the incoherence.

1. Introduction

The noiselet basis, originally described in [2], has garnered interest recently because noiselets (1) are maximally incoherent to the Haar basis and (2) have a fast algorithm for their implementation. Thus, they have been employed in compressive sampling to sample signals that are sparse in the Haar domain [1].

The work presented here was motivated by the observation that it had not been previously shown in a straightforward way that the discrete Haar transform is maximally incoherent to a discretized version of the noiselet transform. Additionally, the exact form of a noiselet matrix needed to be inferred from the original work.

The main contributions are the derivation of a recursive, tensor product-based, construction of noiselet matrices, the unitary matrices that result from the noiselet transform for discrete input, and an intuitive proof showing its incoherence to the corresponding Haar matrix.

2. Preliminaries

2.1 General definitions

Definition 1. *Let A be an $m \times n$ matrix, and B be a matrix of an arbitrary size. The Kronecker product of A and B is*

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}.$$

The Kronecker product (see e.g. [4]) is a bilinear and associative operator which is not generally commutative. It can be combined with a standard maxtrix multiplication as follows:

$$(A \otimes B)(C \otimes D) = AC \otimes BD$$

whenever the products AC, BD exist. This property is sometimes called the *mixed product property*.

Definition 2. *Let A be a $m \times n$ matrix. $A(k,*)$ denotes the (row) vector $(A(k,1) \quad A(k,2) \quad \dots \quad A(k,n))$ while, $A(*,l)$ similarly denotes the (column) vector $(A(1,l) \quad A(2,l) \quad \dots \quad A(m,l))^T$.*

2.2 Noiselets

Noiselets [2] are functions that are completely uncompressible under the Haar transform. The family of noiselets is constructed on the interval $[0,1)$ as follows:

$$\begin{aligned} f_1(x) &= \chi_{[0,1)}(x), \\ f_{2n}(x) &= (1-i)f_n(2x) + (1+i)f_n(2x-1) \\ f_{2n+1}(x) &= (1+i)f_n(2x) + (1-i)f_n(2x-1) \end{aligned}$$

Here, $\chi_{[0,1)}(x) = 1$ on the definition interval $[0,1)$ and 0 otherwise. It is shown in [2] that $\{f_j\}$ is a basis:

Theorem 1. *The set $\{f_j | j = 2^N, \dots, 2^{N+1}-1\}$ is an orthogonal basis of the vector space V_{2^N}, which is the space of all possible approximations at the resolution 2^N of functions in $L^2[0,1)$.*

2.3 Haar Transform

Haar wavelet transform can be described by a real square matrix. For our purposes, it is advantageous to recursively build the Haar matrix using the Kronecker product [3]:

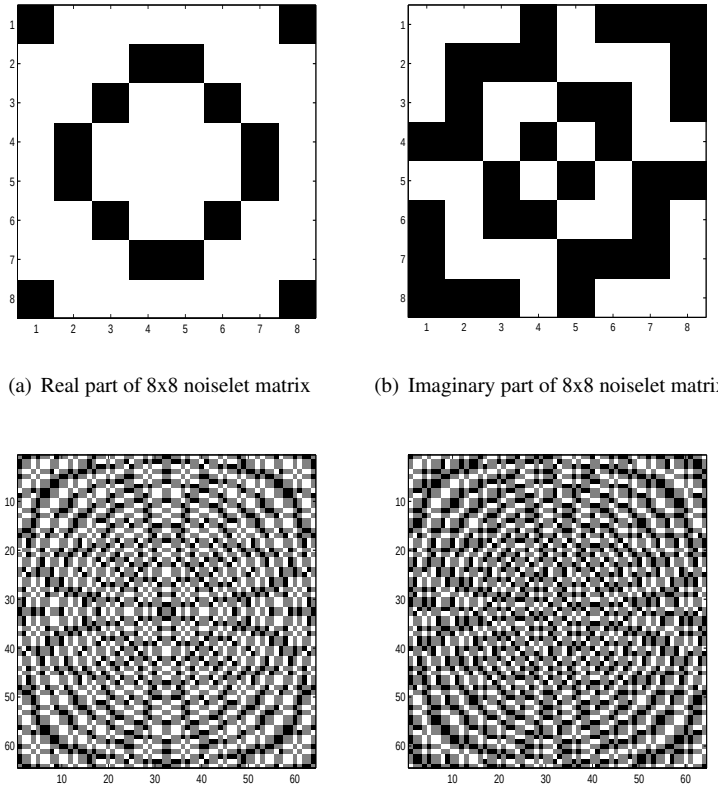
$$H_n = \frac{1}{\sqrt{2}} \begin{bmatrix} H_{n/2} \otimes (1 \quad 1) \\ I_{n/2} \otimes (1 \quad -1) \end{bmatrix}.$$

The iteration starts with $H_1 = \begin{bmatrix} 1 \end{bmatrix}$. The normalization constant $\frac{1}{\sqrt{2}}$ ensures that $H_n^T H_n = I$. Haar wavelets are the rows of H_n.

3. Matrix construction of noiselets

First we extend and discretize the noiselet functions.

Definition 3. *The extensions of noiselets to the interval $[0, 2^m - 1]$ sampled at points $0, \dots, 2^m - 1$ is the series*



(a) Real part of 8x8 noiselet matrix (b) Imaginary part of 8x8 noiselet matrix

(c) Real part of 64x64 noiselet matrix (d) Imaginary part of 64x64 noiselet matrix

Figure 1: Noiselet matrix: graphical view. In figures (a) and (b), the black and white colors denote values of -0.25 and 0.25 respectively. In figures (c) and (d), the black, gray and white colors denote values of -0.125, 0 and 0.125 respectively. .

of functions $f_m(k,l)$

$$f_m(1,l) = \begin{cases} 1 & l = 0,\ldots,2^m-1 \\ 0 & \textit{otherwise} \end{cases}$$
$$f_m(2k,l) = (1-i)f_m(k,2l) + (1+i)f_m(k,2l-2^m)$$
$$f_m(2k+1,l) = (1+i)f_m(k,2l) + (1-i)f_m(k,2l-2^m)$$

where m *denotes the range of extension,* $k = 1,\ldots,2^{m+1}$ *is the function index and* $l = 0,\ldots,2^m-1$ *is the sample index.*

Starting with a 1×1 matrix N_1, a sequence of noiselet matrices $N_1, N_2, N_4, \ldots, N_{2^m}$ of sizes 1×1, 2×2, 4×4, $\ldots$, $2^m \times 2^m$, respectively, is generated. The rows of the N_n matrix are noiselets which form an orthonormal basis for the space $\mathbb{C}^n$.

Definition 4. *For* $n = 1$, $N_1 = \begin{bmatrix}1\end{bmatrix}$. *Then the* $n\times n$ *noiselet matrix* N_n *is built up recursively according to:*

$$N_n(k,*) = \frac{1}{2}(1-i \quad 1+i) \otimes N_{n/2}(\frac{k}{2},*)$$

when $k = 0,2,4,\ldots,n-2$ *and*

$$N_n(k,*) = \frac{1}{2}(1+i \quad 1-i) \otimes N_{n/2}(\frac{k-1}{2},*)$$

when k=1,3,...,n-1.

Lemma 1. *Let* $m > 0$. *The noiselet matrices* $N_1, N_2, N_4, \ldots, N_{2^m}$ *are built up from a series of discretised and extended noiselets* f_m*:*

$$N_n(k,l) = f_m(n+k, \frac{2^m}{n}l), \quad k,l = 0,\ldots,n-1.$$

Proof. Let $m > 0$ be fixed. For $n = 1$

$$N_1(0,0) = f_m(1,0) = 1.$$

By induction, for a matrix of size $n = 2^p, p = 1,\ldots,m$, its basis vector $k = 0,2,4,\ldots,n-2$ and vector indices $l = 0,\ldots,\frac{n}{2}-1$

$$\begin{aligned} N_n(k,l) &= (1-i)N_{n/2}(\frac{k}{2},l) \\ &= (1-i)f_m(\frac{n}{2}+\frac{k}{2}, \frac{2^m}{\frac{n}{2}}l) = f_m(n+k,\frac{2^m}{n}l). \end{aligned}$$

For the same n,k and $l = \frac{n}{2},\ldots,n-1$,

$$\begin{aligned} N_n(k,l) &= (1+i)N_{n/2}(\frac{k}{2},l-\frac{n}{2}) \\ &= (1+i)f_m(\frac{n}{2}+\frac{k}{2}, 2\frac{2^m l}{n} - 2^m) = f_m(n+k,\frac{2^m}{n}l). \end{aligned}$$

To see this, observe that f_m is zero outside of $[0,2^m-1]$ and therefore, the first half of samples of $f_m(k,l)$ are defined exclusively by the expression $(1\pm i)f_m(k,2l)$

whereas the second half of the samples are defined exclusively by $(1 \pm i) f_m(k, 2l - 2^m)$.
For k odd ($k = 1, 3, \ldots, n-1$) the proof is similar. □

Specially, the noiselet matrix N_n for $n = 2^m$ can be found as the "tail" of the function series f_m. Indeed, the expression in Theorem 1 becomes $N(k,l) = f_m(n+k, l)$ for $n = 2^m$.

4. Incoherence of noiselets and Haar

In what follows, we adhere to the terminology of basis coherence which is common in the field of compressive sampling. See for example [1] for details on these definitions and related literature.
Mutual coherence of two bases is defined as the maximum scalar product of any pair of their basis vectors:

Definition 5. *Mutual coherence between two orthonormal bases Ψ, Φ is*

$$\mu(\Psi, \Phi) = \max_{k,j} |\langle \psi_k, \phi_j \rangle|.$$

The minimal coherence is usually termed *maximal* or *perfect* incoherence, which means that $\mu(\Psi, \Phi) = O(1)$. In other words, the matrix of scalar products $\Psi\Phi^*$ is "flat". As Candès and Romberg suggest [1], we will show the perfect incoherence of Haar and noiselets in the following setting. Given an orthonormal $n \times n$ Haar matrix H, we compute the matrix of scalar products for a corresponding noiselet matrix N normalized such that $N^*N = nI$. By doing so, the product will be flat with all values having the magnitude of 1.
For clarity of the main proof, it saves some technical work to define a "twisted" noiselet basis.

Definition 6. *The twisted noiselet matrix $\hat{N}_1 = \begin{bmatrix}1\end{bmatrix}$.*
Then the $n \times n$ twisted noiselet matrix $\hat{N}_n$ is built up recursively by

$$\hat{N}_n(k, *) = \frac{1}{2} \hat{N}_{n/2}(\frac{k}{2}, *) \otimes (1 - i \quad 1 + i)$$

when $k = 0, 2, 4, \ldots, n-2$ and

$$\hat{N}_n(k, *) = \frac{1}{2} \hat{N}_{n/2}(\frac{k-1}{2}, *) \otimes (1 + i \quad 1 - i)$$

when $k = 1, 3, \ldots, n-1$.

The difference between this and the definition of the noiselet matrix N (Definition 4) is that the order of operands in the Kronecker product is changed. In fact, each one is just a permutation of the other.

Lemma 2. *For $n = 2^m$, the bases N_n, $\hat{N}_n$ consist of the same set of basis vectors.*

Proof. Indeed, we can write $\hat{N}_n = P_n N_n$ where P is the permutation matrix:

$$P(k, *) = \begin{cases} (1 \quad 0) \otimes P_{n/2}(\frac{k}{2}, *) & k = 0, 2, 4, \ldots, n-2 \\ (0 \quad 1) \otimes P_{n/2}(\frac{k-1}{2}, *) & k = 1, 3, \ldots, n-1 \end{cases}$$

starting with $P = [1]$.

The claim holds for $n = 1$. For $n = 2, 4, 8, \ldots, 2^m$,

$$P_n N_n(k, l) = P_n(k, *) N_n(l, *)^T$$

as it can easily be shown that N_n is symmetric. Using the recurrent equations for P_n and N_n and applying the mixed product rule, we get, for $k = 0, 2, 4, \ldots, n-2$,

$$P_n N_n(k, l) = \frac{1}{2}(1 - i) P_{n/2}(\frac{k}{2}, *) N_{n/2}(*, \frac{l}{2})$$

when $l = 0, 2, 4, \ldots, n-2$ and

$$P_n N_n(k, l) = \frac{1}{2}(1 + i) P_{n/2}(\frac{k-1}{2}, *) N_{n/2}(*, \frac{l}{2})$$

when $l = 1, 3, \ldots, n-1$. By induction,

$$P_n N_n(k, *) = \frac{1}{2} \hat{N}_{n/2}(\frac{k}{2}, *) \otimes (1 - i \quad 1 + i)$$

for even k indices. This situation for odd k is similar. □

Now the main result can be shown.

Theorem 2. *Let $n = 2^m$ where m is a non-negative integer. Let N_n be the noiselet matrix of size $n \times n$ and let H_n be the Haar matrix of size $n \times n$. Then H_n and N_n are maximally incoherent.*

Proof. Without loss of generality, assume the bases are normalized such that $H_n^T H_n = I$ and $N_n^* N_n = nI$. For the case of $n = 1$,

$$H_1 N_1^* = \begin{bmatrix}1\end{bmatrix} \cdot \begin{bmatrix}1\end{bmatrix} = \begin{bmatrix}1\end{bmatrix}$$

For $n = 2^m, m > 1$, the incoherence is shown by induction. Suppose we know maximal incoherence holds for $\frac{n}{2}$ and we want to show it for n. In the induction step, we use the iterative construction of the Haar matrix by means of Kronecker product. By computing the product

$$H_n \hat{N}_n^* = H(N_n^* P_n^*) = (H_n N_n^*) P_n^T$$

we will still be able to conclude on magnitude of the elements of $(H_n N_n^*)$, since permutation matrices do not change magnitudes.
The product $H_n \hat{N}_n^*$ can be computed per-column; we take the j-th column of $\hat{N}_n^*$, $j = 0, 2, 4, \ldots, n-2$ and transform it by H_n, getting

$$\begin{aligned} H_n \hat{N}_n^*(*, j) = & \frac{1}{\sqrt{2}} \begin{bmatrix} H_{n/2} \otimes (1 \quad 1) \\ I_{n/2} \otimes (1 \quad -1) \end{bmatrix} \\ & \cdot \frac{1}{\sqrt{2}} \hat{N}_{n/2}^*(*, \frac{j}{2}) \otimes (1 - i \quad 1 + i)^* \end{aligned}$$

Note the altered normalization factor of noiselets. Now the mixed product property can be applied to get

$$\frac{1}{2} \begin{bmatrix} H_{n/2} \hat{N}_{n/2}^*(*, \frac{j}{2}) \otimes (1 \quad 1) \begin{bmatrix} 1+i \\ 1-i \end{bmatrix} \\ I_{n/2} \hat{N}_{n/2}^*(*, \frac{j}{2}) \otimes (1 \quad -1) \begin{bmatrix} 1+i \\ 1-i \end{bmatrix} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} H_{n/2} \hat{N}_{n/2}^*(*, \frac{j}{2}) * 2 \\ I_{n/2} \hat{N}_{n/2}^*(*, \frac{j}{2}) * 2i \end{bmatrix}.$$

By induction, it follows that $|H_{n/2} \hat{N}_{n/2}^*(i, \frac{j}{2})| = 1$ and $|I_{n/2} \hat{N}_{n/2}^*(i, \frac{j}{2})| = 1$ for $i = 1, \ldots, \frac{n}{2}$. The Kronecker multiplication is only by entries with magnitude 2, thus the resulting magnitudes are $\frac{1}{2} * 2 = 1$. The proof is equivalent for $j = 1, 3, \ldots, n-1$. □

References:

[1] Emmanuel Candès and Justin Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23(3):969–985, 2007.

[2] R. Coifman, F. Geshwind, and Y. Meyer. Noiselets. *Applied and Computational Harmonic Analysis*, 10:27–44, 2001.

[3] B.J. Falkowski and S. Rahadja. Walsh-like functions and their relations. In *IEE Proceedings on Vision, Image and Signal Processing*, volume 143, pages 279 – 284, 1996.

[4] Alan J. Laub. *Matrix Analysis for Scientists and Engineers*. SIAM, 2005.

Adaptive compressed image sensing based on wavelet modeling and direct sampling

Shay Deutsch [(1)], Amir Averbuch [(1)] and Shai Dekel [(2)]

(1) Tel Aviv University. Israel
(2) GE Healthcare, Israel
shayseut@post.tau.ac.il, Shai.dekel@ge.com, amir@math.tau.ac.il

Abstract:

We present Adaptive Direct Sampling (ADS), an algorithm for image acquisition and compression which does not require the data to be sampled at its highest resolution. In some cases, our approach simplifies and improves upon the existing methodology of Compressed Sensing (CS), by replacing the 'universal' acquisition of pseudo-random measurements with a direct and fast method of adaptive wavelet coefficient acquisition. The main advantages of this direct approach are that the decoding algorithm is significantly faster and that it allows more control over the compressed image quality, in particular, the sharpness of edges.

1. Introduction

Compressed Sensing (CS) [1, 3, 4, 6] is an approach to simultaneous sensing and compression which provides mathematical tools that, when coupled with specific acquisition hardware architectures, can perhaps reduce the acquired dataset sizes, without reducing the resolution or quality of the compressed signal. CS builds on the work of Candès, Romberg, and Tao [4] and Donoho [6] who showed that a signal having a sparse representation in one basis can be reconstructed from a small number of non-adaptive linear projections onto a second basis that is incoherent with the first. The mathematical framework of CS is as follows:

Consider a signal $x \in \mathbb{R}^N$ that is k-sparse in the basis Ψ for $\mathbb{R}^N$. In terms of matrix representation we have $\Psi x = f$, in which f can be well approximated using only $k << N$ non zero entries and Ψ is called the sparse basis matrix. Consider also an $n \times N$ measurement matrix Φ, where the rows of Φ are incoherent with the columns of Ψ. The CS theory states that such a good approximation of signal x can be reconstructed by taking only $n = O(k \log N)$ linear, non adaptive measurements as follows: [1, 3]:

$$y = \Phi x , \qquad (1.1)$$

where y represents an $n \times 1$ sampled vector. Working under this 'sparsity' assumption an approximation to x can be reconstructed from y by 'sparsity' minimization, such as l_1 minimization

$$\min_{\Phi\Psi^{-1} f = y} \|f\|_{l_1} \qquad (1.2)$$

1.2 The "single pixel" camera

For imaging applications, the CS framework has been applied within a new experimental architecture for a 'single pixel' digital camera [10]. The CS camera replaces the CCD and CMOS acquisition technologies by a ***Digital Micro-mirror Device (DMD)***. The DMD consists of an array of electrostatically actuated micro-mirrors where each mirror of the array is suspended above an individual SRAM cell. In [10] the rows of the CS sampling matrix Φ are a sequence of n pseudo-random binary masks, where each mask is actually a 'scrambled' configuration of the DMD array (see also [2]). Thus, the measurement vector y, is composed of dot-products of the digital image x with pseudo-random masks. At the core of the decoding process, that takes place at the viewing device, there is a minimization algorithm solving (1.2). Once a solution is computed, one obtains from it an approximate 'reconstructed' image by applying the transform Ψ to the coefficients. The CS architecture of [10] has few significant drawbacks:

1. Poor control over the quality of the output compressed image: the CS architecture of [10] is not adaptive and the number of measurements is determined before the acquisition process begins, with no feedback during the acquisition process on the progressive quality.
2. Computationally intensive sampling process: Dense measurement matrices such as the sampling operator of the random binary pattern are not feasible because of the huge space and multiplication time requirements. Note that in the one single pixel camera, the sampling operator is based on the random binary pattern, which requires a huge memory and a high computation cost. For example, to get 512×512 image with 64k measurements (25% sampling rate) a random binary operator requires nearly a gigabyte of storage and Giga-flop operations, which makes the recovery almost impossible [14]. The designing of an efficiently measurement basis was proposed [14, 16] by using highly sparse measurements operators, which solve the infeasibility of Gaussian measurement matrix or a random binary masks such as in the one pixel camera. Note, however, in [16], the trade-off between acquisition time and visual quality. To obtain good visual quality, when using TV minimization (which significantly increase the decoding time, compared to LP decoding time)

recovery times of a 256×256 'boat' image are around 60 min.

3. Computationally intensive reconstruction algorithm: It is known that all the algorithms for the minimization (1.2) are very computationally intensive.

2. Direct and adaptive image sensing

Our proposed architecture aims to overcome the drawbacks of the existing CS approach and achieve the following design goals:

1. An acquisition process that captures n measurements, with $n << N$ and $n = O(k)$, where N is the dimension of the full high-resolution image, assumed to be 'k-sparse'. The acquisition process is allowed to adaptively take more measurements if needed to achieve some compressed image target quality.

2. A decoding process which is not more computationally intensive than the existing algorithm in use today such as JPEG or JPEG2000 decoding.

We now present our ADS approach: Instead of acquiring the visual data using a representation that is incoherent with wavelets, we sample directly in the wavelet domain. We use the DMD array architecture in a very different way than in [10]:

1. Any wavelet coefficient is computed from two measurements of the DMD array.

2. We take advantage of the 'feedback' architecture of the DMD where we make decisions on future measurements based on values of existing measurements. This adaptive sampling process relies on a well-known modeling of image edges using a wavelet coefficient tree-structure and so decisions on which wavelet coefficients should be sampled next are based on the values of wavelet coefficients obtained so far [8, 9]. First we explain how the DMD architecture can be used to calculate a wavelet coefficient from two DMD measurements. Modeling an image as a function $f \in L_2(\mathbb{R}^2)$, we have the wavelet representation $f(x) = \sum_{e,j,l} \langle f, \tilde{\psi}^e_{j,l} \rangle \psi^e_{j,l}$, where $e = 1,2,3$ is the subband, $j \in \mathbb{Z}$ the scale and $l \in \mathbb{Z}^2$ the shift. For orthonormal wavelets $\tilde{\psi}^e_{j,l} = \psi^e_{j,l}$. If we consider the Haar basis as an example, than a bivariate Haar wavelet coefficient of type 1 can be computed as follows

$$\langle f, \psi^1_{j,l} \rangle = 2^j \left(\int_{2^{-j} l_1}^{2^{-j}(l_1+1)} \int_{2^{-j} l_2}^{2^{-j}(l_2+1/2)} f(x_1,x_2)\,dx_1 dx_2 - \int_{2^{-j} l_1}^{2^{-j}(l_1+1)} \int_{2^{-j}(l_2+1/2)}^{2^{-j}(l_2+1)} f(x_1,x_2)\,dx_1 dx_2 \right), \quad (2.1)$$

i.e., the difference of pixel sums over two neighboring dyadic rectangles multiplied by 2^j. By Similar computation we can sample the Haar wavelet coefficients of the second and third kinds with two measurements. Moreover, there exist DMD arrays with micro-mirrors that can produce a grayscale value, not just 0 or 1 (contemporary DMD can produce 1024 grayscale value). We can use these devices for computation of arbitrary wavelet transforms, where the computation of each coefficient requires only two measurements, since the result of any real-valued functional g acting on the data can be computed as a difference of two 'positive' g_+, g_- 'functionals', i.e. ,where the coefficients are positive: $g = g_+ - g_-$, $g_+, g_- \geq 0$.

3. Modeling of image edges by wavelet tree-Structures and the ADS algorithm

Most of the significant wavelet coefficients are located in the vicinity of edges. Wavelets can be regarded as multi-scale local edge detectors, where the absolute value of a wavelet coefficient corresponds to the local strength of the edge. We impose the tree-structure of the wavelet coefficients. Due to the analysis properties of wavelets, coefficient values tend to persist through scale. A large wavelet coefficient in magnitude generally indicates the presence of singularity inside its support. A small wavelet coefficient generally indicates a smooth region. We use this nesting property and acquire wavelet coefficients in the higher resolutions if their parent is found to be significant. For further detection of singularities at fine scales, we estimate the Lipschitz exponent.

3.1 The Lipschitz exponent

Our goal is to estimate the significance of wavelet coefficients that were not sampled yet, using values of coefficients that were already sampled. To this end we use the well known characterization of local Lipschitz smoothness by the decay of wavelet coefficients across scales [12]. A function f is said to be Lipschitz α in the neighborhood of (x_1, x_2) if there exists ε_1 and ε_2 as well as $A > 0$ such that for any $h_1 < \varepsilon_1$ and $h_2 < \varepsilon_2$

$$|f(x_1+h_1, x_2+h_2) - f(x_1,x_2)| \leq A(h_1^2 + h_2^2)^{\alpha/2} \quad (3.1)$$

We actually use a subtler, 'directional' notion of local Lipschitz smoothness. So, for example, for the horizontal subband, $e = 1$, we defined local α_1 Horizontal Lipschitz smoothness by the minimal $A > 0$ satisfying for $h_1 < \varepsilon_1$

$$|f(x_1+h_1, x_2) - f(x_1,x_2)| \leq A h_1^{\alpha_1}.$$

If the function is locally α_e Lipschitz at (x_1, x_2) then for any wavelet $\tilde{\psi}^e_{j,l}$ whose support contains (x_1, x_2), we have that $|\langle f, \tilde{\psi}^e_{j,l} \rangle| \leq C(2^j)^{\alpha_e}$. By taking the logarithm we have

$$\log_2 |\langle f, \tilde{\psi}^e_{j,l} \rangle| \leq \alpha j + \log_2(C). \quad (3.3)$$

Thus the Lipschitz exponents can be determined from the slope of the decay of $\log_2 \left| \left\langle f, \tilde{\psi}^e_{j,l} \right\rangle \right|$ across scales (see also [15]). These slopes are considered measurements of local singularities, such that when $0 < \alpha_e < 1$ a function f has a directional singularity which increases as $\alpha_e \to 0$. Thus we estimate the existence of local directional singularities and the significance of unsampled coefficients at high scales, using estimates of local directional Lipschitz exponents from wavelet coefficients that were already sampled.

3.2 The ADS Algorithm

Our adaptive CS algorithm works as follows:

1. Acquire the values of all low-resolution coefficients up to a certain low-resolution J. Each computation is done using two DMD array measurements as in (2.1). In one embodiment the initial resolution J can be selected as $\left\lfloor \frac{\log_2 N}{2} \right\rfloor + const$. In any case, J should be bigger if the image is bigger. Note that the total number of coefficients at resolutions $\geq J$ is $2^{2(1-J)} N$, which is a small fraction of N.

2. Initialize a 'sampling queue' containing the indices of each of the four children of significant coefficients at the resolution J. Thus for a significant coefficient with index (e, J, l), we add to the queue the coefficients with indices: $(e, J-1, (2l_1, 2l_2))$, $(e, J-1, (2l_1, 2l_2+1))$, $(e, J-1, (2l_1+1, 2l_2))$ and $(e, J-1, (2l_1+1, 2l_2+1))$.

3. Process the sampling queue until it is exhausted as follows:

a. Sample the wavelet coefficient corresponding to the index (e, j, l) at the beginning of the queue using two DMD array measurements (see Section 2).

b. Add to the end of the queue the indices of the coefficient's four children, only if one of the following holds:

(i) The coefficient is at a resolution $j > J-2$ and the coefficient's absolute value is greater than a given threshold t_{low}.

(ii) The coefficient is at resolution $1 < j \leq J-2$ and the corresponding estimated absolute value of its children using the local Lipschitz exponent method (see Section 3.1) is greater than a given threshold t_{high}.

c. Remove the processed index from the queue and go to step (a).

In a way, our algorithm can be regarded as an adaptive edge acquisition device where the acquisition resolution increases only in the vicinity of edges! Observe that the algorithm is output sensitive. Its time complexity is of the order n where n is the total number of computed wavelet coefficients, which can be substantially smaller than the number of pixels N. The number of samples is influenced by the size of the thresholds used by the algorithm in step 3.b. It is also important to understand that the number of samples is influenced by the amount of visual activity in the image. If there are more significant edges in the image, then their detection at lower resolutions will lead to adding higher resolution sampling to the queue.

4. Experimental results

To evaluate our approach, we use the optimal k-term wavelet approximation as a benchmark. It is well known [5] that for a given image with N pixels, the optimal orthonormal wavelet approximation using only k coefficients is obtained using the k largest coefficients

$$\left| \left\langle f, \psi^{e_1}_{j_1,l_1} \right\rangle \right| \geq \left| \left\langle f, \psi^{e_2}_{j_2,l_2} \right\rangle \right| \geq \left| \left\langle f, \psi^{e_3}_{j_3,l_3} \right\rangle \right| \geq \cdots,$$

$$\left\| f - \sum_{i=1}^{k} \left\langle f, \psi^{e_i}_{j_i,l_i} \right\rangle \psi^{e_i}_{j_i,l_i} \right\|_{L_2(\mathbb{R}^2)} = \min_{\#\Lambda = k} \left\| f - \sum_{(e,j,l) \in \Lambda} \left\langle f, \psi^{e}_{j,l} \right\rangle \psi^{e}_{j,l} \right\|_{L_2(\mathbb{R}^2)}.$$

For biorthogonal wavelets this 'greedy' approach gives a near-best result, i.e. within a constant factor of the optimal k-term approximation. One can apply thresholding and construct a k-term approximation using only coefficients whose absolute value is above the threshold, which still requires the order of N computations. In contrast, our ADS algorithm is output sensitive and requires only order of n computations. To simulate our algorithm in software, we first pre-compute the entire wavelet transform of a given image. However, we strictly follow the recipe of our ADS algorithm and extract a wavelet coefficient from the pre-computed coefficient matrix only if its index was added to the adaptive sampling queue. In fig 1(a) we see a 'benchmark' near-best 7000-term biorthogonal [9,7] wavelet approximation of the Lena image, extracted from the 'full' wavelet representation by thresholding. In fig 1(b) we see a 6782-term approximation extracted from an ADS adaptive sampling process with n=12796 sampled wavelet coefficient.

(a) 7000-term

(b) ADS 6782-term

Fig.1. (a) Near-best 7000-term [9,7] approximation computed from the 'full' wavelet representation N=262,144, PSNR=31 dB (b) ADS 6782-term [9,7] approximation, extracted from n=12,796 adaptive wavelet samples, PSNR=28.7 dB.

5. Conclusion

We present an architecture that acquires and compresses high resolution visual data, without fully sampling the entire data at its highest resolution. By sampling in the wavelet domain we are able to acquire low resolution coefficients within a small number of measurements. We then exploit the wavelet tree structure to build an adaptive sampling process of the detail wavelet coefficients. Experimental results show good visual and PSNR results with a small number of measurements. The coefficients acquired by the ADS algorithm can be streamed into a tree-based wavelet compression algorithm whose decoding time is significantly faster then the solution of (1.2).

REFERENCES

1. R. Baraniuk, Compressive Sensing, Lecture Notes in IEEE Signal Processing Magazine, Vol. 24, No. 4, pp. 118-120, July 2007.
2. R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, A simple proof of the restricted isometry property for random matrices, Constructive Approximation 28 (2008), 253-263.
3. E. Candès, Compressive sampling, Proc. International Congress of Mathematics, 3 (2006), 1433-1452.
4. E. Candès, J. Romberg, and T. Tao, Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information, IEEE Trans. Inf. Theory 52 (2006), 489–509.
5. R. DeVore, Nonlinear approximation, Acta Numerica7 (1998), 50-51.
6. D. Donoho, Compressed sensing, IEEE Trans. Information Theory, 52 (2006), 1289-1306.
7. C. La and M. Do, Signal reconstruction using sparse tree representations, Proc. SPIE Wavelets XI, San Diego, September 2005.
8. A. Said and W. Pearlman, A new fast and efficient image codec based on set partitioning in hierarchical trees, IEEE Trans. Circuits Syst. Video Tech., 6 (1996), 243-250.
9. J. Shapiro, Embedded image coding using zerotrees of wavelet coefficients, IEEE Trans. Signal Process. 41 (1993), 3445-3462.
10 .D. Takhar, J. Laska, M. Wakin, M. Duarte, D. Baron, S. Sarvotham, K. Kelly and R. Baraniuk, A New Compressive Imaging Camera Architecture using Optical-Domain Compression, Proc. of Computational Imaging IV , SPIE, 2006.
11. S. Dekel, Adaptive compressed image sensing based on wavelet-trees, report 2008.
12. S. Mallat, "a wavelet tour of signal processing".
13. S. Mallat and W. L. Hwang, "singularity detection and processing with wavelets," IEEE Trans. Inf. Theory 38,617-642(1992).
14. L. Gan, T. Do, T. Tran , Fast compressive imaging using scrambled Hadamard transform ensemble, preprint 2008.
15. Z. Chen and M. A. Karim, Forest representation of wavelet transforms and feature detection, Opt. Eng. 39 (2000), 1194–1202.
16. R. Berinde, P. Indik, sparse recovery using sparse random matrices, Tech. Report of MIT 2008.
17. F Rooms, A. Pizurica and, W. Philips, estimating image blur in the wavelet domain, IEEE Benelux Signal Processing Symposium (SPS-2002).

Asymmetric Multi-channel Sampling in Shift Invariant Spaces

Sinuk Kang (1) and K.H. Kwon (1)

(1) KAIST, 335 Gwahangro, Yuseong-gu, Daejeon 305-701, S. Korea.
sukang@kaist.ac.kr, khkwon@kaist.edu

Abstract:

We consider a multi-channel sampling with asymmetric sampling rates in shift invariant spaces, while related previous works have supposed that each channel has a symmetric(uniform) sampling rate. Motivated by the fact that shift invariant spaces are isomorphic images of $L^2[0, 2\pi]$, we obtain a sampling expansion in shift invariant spaces by using frame or Riesz basis expansion in $L^2[0, 2\pi]$. The samples in the expansion are expressed in terms of frame coefficients of an appropriate function with respect to a certain frame in $L^2[0, 2\pi]$. The involved reconstruction functions are given explicitly by using the frame operator. We also present relation between asymmetric multi-channel sampling and symmetric one.

1. Introduction

Reconstructing a band-limited signal f from samples which are taken from several channeled versions of f is called multi-channel sampling. The multi-channel sampling method goes back to the work of Shannon [6] and Fogel [2], where the reconstruction of a band-limited signal from samples of the signal and of its derivatives was suggested. Generalized sampling expansion for arbitrary multi-channel sampling was introduced first by Papoulis [5].

Papoulis' result has been extended to a general shift-invariant space [1, 7, 8]. Here, a shift invariant space $V(\phi)$ with a generator $\phi \in L^2(\mathbb{R})$ is defined by the closed subspace of $L^2(\mathbb{R})$ spanned by integer translates $\{\phi(t-n) : n \in \mathbb{Z}\}$ of ϕ. Recently García and Pérez-Villalón [3] derived stable generalized sampling in a shift-invariant space by using some special dual frames in $L^2[0, 1]$.

The previous works related to the multi-channel sampling have assumed that numbers of samples from each channel are uniform, namely, sampling rates of channels are same. In this paper we consider a multi-channel sampling with asymmetric sampling rates in shift invariant spaces. We find an expression for the samples as frame coefficients of an appropriate function in $L^2[0, 2\pi]$ with respect to some particular frame in $L^2[0, 2\pi]$ and present the sufficient and necessary condition under which a sequence of functions of particular form becomes a frame or a Riesz basis for $L^2[0, 2\pi]$. Using isomorphism between a shift invariant space $V(\phi)$ and $L^2[0, 2\pi]$, we derive sampling theory in $V(\phi)$ with some Riesz generator ϕ and find a formula of reconstruction functions by means of the frame operator. The theory contains both a frame and Riesz basis expansion as sampling formulas.

2. Asymmetric multi-channel sampling

Assume that $\phi(t)$ is everywhere well defined complex valued square integrable function on $\mathbb{R}$ throughout the paper. Moreover, let $\phi(t)$ be a Riesz generator with $C_\phi(t) < \infty$ for any $t \in \mathbb{R}$ so that $V(\phi)$ is an RKHS(see Proposition 2.4 in [4]). We now are given a LTI system $\{L_j[\cdot]\}_{j=1}^N$ whose impulse response is $\{l_j(t) : l_j \in L^2(\mathbb{R}),\ j = 1, 2, \cdots, N\}$. The aim of this paper is to recover any $f(t) \in V(\phi)$ via discrete samples from $\{L_j[f]\}_{j=1}^N$ as

$$f(t) = \sum_{j=1}^{N} \sum_{n\in\mathbb{Z}} L_j[f](\sigma_j + r_j n) s_{j,n}(t), \qquad (1)$$

where $\{s_{j,n}(t) : j = 1, \cdots, N \text{ and } n \in \mathbb{Z}\}$ is a frame or a Riesz bases of $V(\phi)$ and $0 \le \sigma_j < r_j$ with a positive integer r_j for $j \in \{1, 2, \cdots, N\}$.

2.1 An expression for the samples

Define an isomorphism J from $L^2[0, 2\pi]$ onto $V(\phi)$ by

$$JF(t) = \frac{1}{2\pi} \sum_{n\in\pi} \langle F(\xi), e^{-in\xi}\rangle \phi(t-n),\ F(\xi) \in L^2[0, 2\pi].$$

By the isomorphism $J : L^2[0, 2\pi] \to V(\phi)$, the reconstruction formula (1) is equivalent to the following one:

$$F(\xi) = \sum_{j=1}^{N} \sum_{n\in\mathbb{Z}} L_j[f](\sigma_j + r_j n) S_{j,n}(\xi),\ F(\xi) \in L^2[0, 2\pi], \qquad (2)$$

where $f(t) = JF(t)$ and $s_{j,n}(t) = JS_{j,n}(t)$. Notice further that $L_j f(\sigma_j + r_j n)$ is represented by an inner product of $F(\xi)$ and some function in $L^2[0, 2\pi]$.

Lemma 2.1.1 *Let $L[\cdot]$ be a LTI system with an impulse response $l(t) \in L^2(\mathbb{R})$ and $\psi(t) = L[\phi](t) = (\phi * l)(t)$.*

(a) L is a bounded operator from $L^2(\mathbb{R})$ into $L^\infty(\mathbb{R})$, $\|f * l\|_\infty \le \|f\|_2 \|l\|_2$ and $Lf(t) \in C_\infty^o(\mathbb{R})$,

(b) $\sup_{\mathbb{R}} C_\psi(t) < \infty$,

(c) *(cf. Lemma 2 in [3]) for any* $f(t) = (\mathbf{c} * \phi)(t)$ *with* $\mathbf{c} \in \ell^2$ *in* $V(\phi)$, $L[f](t) = (\mathbf{c} * \psi)(t)$ *converges absolutely and uniformly on* $\mathbb{R}$. *For any* $f(t) = JF(t) \in V(\phi)$ *with* $F(\xi) \in L^2[0, 2\pi]$,

$$L[f](t) = \langle F(\xi), \frac{1}{2\pi}\overline{Z_\psi(t, \xi)}\rangle_{L^2[0,2\pi]}.$$

In particular,

$$L[f](\sigma_j + r_j n) = \langle F(\xi), \frac{1}{2\pi}\overline{Z_\psi(\sigma_j, \xi)}e^{-ir_j n\xi}\rangle_{L^2[0,2\pi]}. \quad (3)$$

2.2 The sampling theorem

For a given LTI system $\{L_j[\cdot]\}_{j=1}^N$, let $L_j\phi(t) = \psi_j(t)$, $1 \le j \le N$. Using equation (3), the expansion (2) is equivalent to

$$F(\xi) = \sum_{j=1}^{N}\sum_{n\in\mathbb{Z}} \langle F(\xi), \frac{1}{2\pi}\overline{Z_{\psi_j}(\sigma_j, \xi)}e^{-ir_j n\xi}\rangle_{L^2[0,2\pi]} \cdot S_{j,n}(\xi),\ F(\xi) \in L^2[0, 2\pi],$$

where $f(t) = JF(t)$ and $s_{j,n}(t) = JS_{j,n}(t)$.
For convenience, we introduce a few more notations. Let $g_j(\xi) \in L^2[0, 2\pi]$ for $1 \le j \le N$, $g_{j,m_j}(\xi) := g_j(\xi)e^{ir_j(m_j-1)\xi}$ for $1 \le m_j \le \frac{r}{r_j}$ and

$$G(\xi) = [Dg_{1,1}(\xi), Dg_{1,2}(\xi), \cdots, Dg_{1,\frac{r}{r_1}}(\xi), Dg_{2,1}(\xi), \cdots, Dg_{N,\frac{r}{r_N}}(\xi)]^T,$$

where D is an unitary operator from $L^2[0, 2\pi]$ onto $L^2(I)^r$ defined by $(DF)(\xi) = [F(\xi), F(\xi + \frac{2\pi}{r}), \cdots, F(\xi + (r-1)\frac{2\pi}{r})]^T$, $F(\xi) \in L^2[0, 2\pi]$. Note that $G(\xi)$ is the $\sum_{j=1}^N \frac{r}{r_j} \times r$ matrix whose entries are in $L^2[0, \frac{2\pi}{r}]$. And define $\lambda_M(\xi)$(resp. $\lambda_m(\xi)$) as the largest(resp. the smallest) eigenvalue of $r \times r$ matrix $G(\xi)^*G(\xi)$, β_G as $\|\lambda_M(\xi)\|_\infty$ and α_G as $\|\lambda_m\|_0$.

Lemma 2.2.1 *Let* $g_j \in L^2[0, 2\pi]$ *and* r_j *be a positive integer for* $1 \le j \le N$. *Define* r *as the least common multiplier of* $\{r_j\}_{j=1}^N$. *Then* $\{\overline{g_j(\xi)}e^{-ir_j n\xi} : 1 \le j \le N,\ n \in \mathbb{Z}\}$ *is a*

(a) *Bessel sequence in* $L^2[0, 2\pi]$ *if and only if* $\|\lambda_M(\xi)\|_\infty < \infty$ *if and only if* $g_j \in L^\infty[0, 2\pi]$ *for* $1 \le j \le N$. *In this case, optimal bound is* $\frac{2\pi}{r}\|\lambda_M(\xi)\|_\infty$;

(b) *frame of* $L^2[0, 2\pi]$ *if and only if* $0 < \|\lambda_m(\xi)\|_0 \le \|\lambda_M(\xi)\|_\infty < \infty$ *so that* $r \le \sum_{j=1}^N \frac{r}{r_j}$ *and optimal bounds are* $\frac{2\pi}{r}\|\lambda_m(\xi)\|_0 \le \frac{2\pi}{r}\|\lambda_M(\xi)\|_\infty$;

(c) *Riesz basis of* $L^2[0, 2\pi]$ *if and only if frame of* $L^2[0, 2\pi]$ *and* $r = \sum_{j=1}^N \frac{r}{r_j}$, *i.e.,* $1 = \sum_{j=1}^N \frac{1}{r_j}$ *if and only if* $g_j(\xi) \in L^\infty[0, 2\pi]$ *for* $1 \le j \le N$, $1 = \sum_{j=1}^N \frac{1}{r_j}$ *and* $|\det G(\xi)| \ge \exists\alpha > 0$ *a.e.*.

Appealing to the setting $g_j(\xi) = \frac{1}{2\pi}Z_{\psi_j}(\sigma_j, \xi)$ for $1 \le j \le N$, we have

Theorem 2.2.2 *Let* $\phi(t)$ *be a Riesz generator with* $C_\phi(t) < \infty$, $t \in \mathbb{R}$ *and* $\{L_j[\cdot]\}_{j=1}^N$ *be LTI systems with an impulse response* $\{l_j(t)\}_{j=1}^N \in L^2(\mathbb{R})$. *Let* $\{\psi_j(t) = (\phi * l_j)(t)\}_{j=1}^N$, $r_j \ge 1$ *an integer and* $0 \le \sigma_j < r_j$.

(a) *If* $0 < \alpha_G \le \beta_G < \infty$, *i.e.,* $0 < \alpha_G$ *and* $Z_{\psi_j}(\sigma_j, \xi) \in L^\infty[0, 2\pi]$, $1 \le j \le N$, *then there is a frame* $\{s_{j,n}(t) : 1 \le j \le N,\ n \in \mathbb{Z}\}$ *of* $V(\phi)$ *for which*

$$f(t) = \sum_{j=1}^{N}\sum_{n\in\mathbb{Z}} L_j f(\sigma_j + r_j n)s_{j,n}(t),\ f(t) \in V(\phi). \quad (4)$$

(b) *Assume that* $Z_{\psi_j}(\sigma_j, \xi) \in L^\infty[0, 2\pi]$, $1 \le j \le N$. *Then there is a frame* $\{s_{j,n}(t) : 1 \le j \le N,\ n \in \mathbb{Z}\}$ *of* $V(\phi)$ *for which (4) holds if and only if* $0 < \alpha_G$.

(c) *Assume that* $Z_{\psi_j}(\sigma_j, \xi) \in L^\infty[0, 2\pi]$, $1 \le j \le N$. *Then there is a Riesz basis* $\{s_{j,n}(t) : 1 \le j \le N,\ n \in \mathbb{Z}\}$ *of* $V(\phi)$ *for which (4) holds if and only if* $0 < \alpha_G$ *and* $1 = \sum_{j=1}^N \frac{1}{r_j}$.

In all cases, sampling series (4) converges in $L^2(\mathbb{R})$, *absolutely on* $\mathbb{R}$ *and uniformly on any subset of* $\mathbb{R}$ *on which* $C_\phi(t)$ *is bounded.*

Remark 2.2.3 *Asymmetric multi-channel sampling series with LTI system* $\{L_j[\cdot]\}_{j=1}^N$ *whose impulse response is* $\{l_j(t)\}_{j=1}^N$ *can be considered as symmetric multi-channel sampling series with LTI system* $\{\tilde{L}_{j,m_j}[\cdot]\}_{j=1,m_j=1}^{N,\frac{r}{r_j}}$ *with impulse response* $\{\tilde{l}_{j,m_j}(t)\}_{j=1,m_j=1}^{N,\frac{r}{r_j}}$, *where* $\tilde{l}_{j,m_j}(t) = l_j(r_j(m_j - 1) + t)$.

2.3 Reconstruction functions

Let S be a frame operator with frame $\{\overline{g_j(\xi)}e^{-ir_j n\xi}\}_{j,n}$. For any $F(\xi) \in L^2[0, 2\pi]$,

$$SF(\xi) = \sum_{j=1}^{N}\sum_{m_j=1}^{\frac{r}{r_j}} \overline{g_j(\xi)}e^{-ir_j(m_j-1)\xi} \cdot \frac{2\pi}{r}g_{j,m}(\xi)^T DF(\xi)$$

so that

$$DSF(\xi) = \frac{2\pi}{r}G^*G(\xi)DF(\xi).$$

Then, from Lemma 2.2.1 (b), there exists $(G^*G)^{-1}(\xi)$ a.e. such that

$$D(S^{-1}(\overline{g_j(\xi)}e^{-ir_j n\xi})) = \frac{r}{2\pi}(G^*G)^{-1}(\xi)D(\overline{g_j(\xi)}e^{-ir_j n\xi})$$

for $1 \le j \le N$ and $n \in \mathbb{Z}$. Hence,

$$\{s_{j,n}\}_{j,n} = \{\frac{r}{2\pi}JD^{-1}[(G^*G)^{-1}(\xi)D(\overline{g_j(\xi)}e^{-ir_j n\xi})]\}_{j,n}.$$

Remark 2.3.1 *One sufficient condition under which* $\{s_{j,n}\}_{j,n}$ *is translates of a single function in* $L^2[0, 2\pi]$ *is that* r *divides* r_j *for all* $1 \le j \le N$. *Since* r *is the least common multiplier of* $\{r_j\}_{j=1}^N$, *the condition holds if and only if* $r = r_j$ *for all* $1 \le j \le N$.

References:

[1] I. Djokovic, P. P. Vaidyanathan, Generalized sampling theorems in multiresolution subspaces, *IEEE Trans. Signal Process.*, 45:583-599, 1997.

[2] L. J. Fogel, A note on the sampling theorem, *IRE Tran. Infor. Theory IT-1*:47-48, 1995.

[3] A. G. García and G. Pérez-Villarón, Dual frames in $L^2(0,1)$ connected with generalized sampling in shift-invariant spaces, *Appl. Comput. Harmon. Anal.*, 20:422-433, 2006.

[4] J. M. Kim, K. H. Kwon, Sampling expansion in shift invariant spaces, *Intern. J Wavelets, Multiresolution and Inform. Processing*, 6(2):223-248, 2008.

[5] A. Papoulis, Generalized sampling expansion, *IEEE Trans. Circuits Systems*, 24(11), 652-654, 1977.

[6] C. E. Shannon, Communication in the presence of noise, *Proc. IRE*, 37:10-21, 1949.

[7] M. Unser, J. Zerubia, Generalized sampling: Stability and performance analysis, *IEEE trans. Signal Process.*, 45(12):2941-2950, 1997.

[8] M. Unser, J. Zerubia, A generalized sampling theory without band-limiting constraints, *IEEE Trans. Circuits Syst. 2*, 45(8):959-969, 1998.

Sparse Data Representation on the Sphere using the Easy Path Wavelet Transform

Gerlind Plonka [(1)] and Daniela Roşca [(2)]

(1) Department of Mathematics, University of Duisburg-Essen, Campus Duisburg, 47048 Duisburg, Germany.
(2) Department of Mathematics, Technical University of Cluj-Napoca, 400020 Cluj-Napoca, Romania.
gerlind.plonka@uni-due.de, Daniela.Rosca@math.utcluj.ro

Abstract:

In this paper we consider the Easy Path Wavelet Transform (EPWT) on spherical triangulations. The EPWT has been introduced in [7] in order to obtain sparse image representations. It is a locally adaptive transform that works along pathways through the array of function values and exploits the local correlations of the data in a simple appropriate manner. In our approach the usual one-dimensional discrete wavelet transform (DWT), orthogonal or biorthogonal, can be applied.

1. Introduction

One important problem in data analysis is to construct efficient low-level representations using only a very small part of the original data. However, these sparse approximations should provide a precise characterization of relevant features of the data like discontinuities (edges) and texture components.

It is well-known that wavelets can represent piecewise smooth signals efficiently. However, higher-dimensional structures may not be represented suitably by sparse wavelet decompositions based on tensor product wavelets, because directional geometrical properties of the data cannot be adapted.

The last years have seen many attempts to construct locally adaptive wavelet-based schemes that take into account the special geometry of the data. In particular, for sparse representation of images, different ideas, that try to exploit the local correlations of the data, have been developed (see e.g. [1, 2, 3, 4, 5, 6, 7, 10]).

We will focus on the EPWT recently introduced in [7] for sparse image representation. In this paper, we want to adapt the EPWT to triangulations of the sphere.

For this purpose, we apply the idea used by Roşca [8, 9] to obtain a suitable spherical triangulation. We employ a polyhedral subdivision domain. The triangular faces of the polyhedron are successively subdivided into four smaller triangles. Each triangle can be transported radially to the sphere. This approach has been used in [8, 9] for the construction of Haar wavelets and of locally supported rational spline wavelets on the sphere.

The idea of the EPWT on spherical triangulations is very simple. First we fix a certain neighborhood of a triangle, e.g. the three triangles that have common edges with the reference triangle. Next, we use a one-dimensional indexing of all triangles of the fixed triangulation and assume that each function value of a given data vector is associated to one triangle, or rather to its corresponding (one-dimensional) index.

In the first step we select a path through the complete index set in such a way that data points associated to neighbor indices in the path are strongly correlated. For this purpose, for each index we choose "the best" neighbor index that has not been used in the path yet, such that the absolute difference between neighboring data values is the smallest. The complete path vector can be seen as a permutation of the original index vector. Then we apply a suitable (one-dimensional) discrete wavelet transform to the data vector along the path, and the choice of the path will ensure that most wavelet coefficients remain small. The same procedure can be successively applied to the down-sampled data. After a suitable number of iterations, we apply a shrinkage procedure to all wavelet coefficients in order to find a sparse digital representation of the function. For reconstruction one needs the path vector at each level in order to apply the inverse wavelet transform.

2. Spatial and spherical triangulations

Consider the sphere $\mathbb{S}^2 = \{\mathbf{x} \in \mathbb{R}^3,\ \|\mathbf{x}\|_2 = 1\}$ and let Π be a convex polyhedron with triangular faces, containing O inside. For example we can take an icosahedron, a cube with triangulated faces, an octahedron, etc. The boundary of the polyhedron will be denoted by Ω. We denote by $\mathcal{T}^0 = \{T_1, \ldots, T_M\}$ the set of faces of Π. For each triangle $T \in \mathcal{T}^0$ we take the mid-points of its edges and construct four triangles of equal area, as in Figure 1. All these small triangles will form a refined triangulation of $\mathcal{T}^0$, denoted $\mathcal{T}^1$. Continuing the refinement process in the same manner, we obtain a triangulation $\mathcal{T}^j$ of Ω, for $j \in \mathbb{N}$. For application of the EPWT we will stop the refinement process at a suitable sufficiently high (fixed) level j depending on the data set in the application. For application of the EPWT we will need a one-dimensional index set $J = J^j$ for the triangles in $\mathcal{T}^j$. Using the octahedron, this one-dimensional index set J can be as in Figure 1 (right). Observe that for the octahedron the number of triangles at the jth level is given by $\#J = \#\mathcal{T}^j = 2^{2j+3}$.

In order to obtain a spherical triangulation, for the given

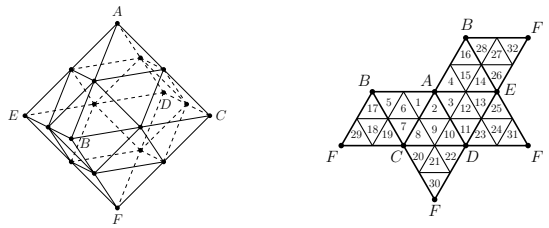


Figure 1: Illustration of the octahedron with triangulation $\mathcal{T}^1$ (left) and a fold apart version of the octahedron on the plane, with a one-dimensional indexing of all triangles.

polyhedron Π we define the radial projection $p : \Omega \to \mathbb{S}^2$,

$$p(x,y,z) = (x^2+y^2+z^2)^{-1/2} \cdot (x,y,z), \quad (x,y,z) \in \Omega.$$

The set $\mathcal{U}^j = \{U = p(T),\, T \in \mathcal{T}^j\}$ will be a triangulation of the sphere $\mathbb{S}^2$. For indexing the spherical triangles in $\mathcal{U}^j$, we use the same index set J as for the triangulation $\mathcal{T}^j$ of the polyhedron.

3. Definitions and Notations for the EPWT

In order to explain the idea of the EPWT, where we want to use the discrete one-dimensional wavelet transform along *path vectors* through the data, we need some definitions and notations.

Let us assume that a fixed refined spherical triangulation $\mathcal{U}^j$ is given.Let J be a one-dimensional index set for the spherical triangles in $\mathcal{U}^j$.

We define a *neighborhood* of an index $\nu \in J$ as

$$\mathcal{N}(\nu) = \{\mu \in J \backslash \{\nu\} : T_\mu \text{ and } T_\nu \text{ have a common edge}\}.$$

Hence, each index $\nu \in J$ has exactly three neighbors. One may also use a bigger neighborhood, e.g. $\mathcal{N}(\nu) = \{\mu \in J \setminus \{\nu\} : T_\mu$ and T_ν have a common edge or a common vertex $\}$, in which case each index has 12 neighbors.

We also need a definition of neighborhood of subsets of an index set. We shall consider disjoint *partitions* of J of the form $\{J_1, J_2, \ldots, J_r\}$, where $J_\mu \cap J_\nu = \emptyset$ for $\mu \neq \nu$ and $\bigcup_{\nu=1}^r J_\nu = J$. We then say that two different subsets J_ν and J_μ from the partition are *neighbors*, and we write $J_\nu \in \mathcal{N}(J_\mu)$, if there exist the indices $l \in J_\nu$ and $l_1 \in J_\mu$ such that $l \in \mathcal{N}(l_1)$. We consider a function f being piecewise constant on the triangles of $\mathcal{U}^j$, i.e., we identify each spherical triangle in $\mathcal{U}^j$ with a value of f. Hence, f is uniquely determined by the data vector $(f_\nu)_{\nu \in J}$.

We will look for path vectors through index subsets of J and we apply a one-dimensional wavelet transform along these path vectors. Any orthogonal or biorthogonal one-dimensional wavelet transform can be used here.

4. Description of the EPWT

In this section we give a summary of the idea of the EPWT, described in more details in [7]. We start with the decomposition of the real data $(f_\nu)_{\nu \in J}$, and we assume that $N = \#J$ is a multiple of 2^L with $L \in \mathbb{N}$. Then we will be able to apply L levels of the EPWT. For the considered octahedron we have $N = 2^{2j+3}$.

Decomposition

First level

We first determine a complete path vector $\mathbf{p}^L$ through the index set $J = \{1, 2, \ldots, N\}$ and then apply a suitable discrete one-dimensional (periodic) wavelet transform to the function values $\mathbf{f}^L = (\mathbf{f}^L(j))_{j \in J}$ along the path $\mathbf{p}^L$. We start with $\mathbf{p}^L(1) := 1$. Next, for $\mathbf{p}^L(2)$ we take

$$\mathbf{p}^L(2) := \operatorname*{argmin}_k \{|\mathbf{f}^L(1) - \mathbf{f}^L(k)|,\ k \in \mathcal{N}(1)\}.$$

We proceed in this manner, thereby determining a path vector through the index set J, that is locally adapted to the function f (easy path). With the procedure described above, we obtain a pathway such that the absolute differences between neighboring function values $\mathbf{f}^L(l)$ along the path are as small as possible. In general, for a given the index $\mathbf{p}^L(l)$, $1 \leq l \leq N-1$, the next value $\mathbf{p}^L(l+1)$ is defined by

$$\begin{aligned}\mathbf{p}^L(l+1) \ := \ & \operatorname*{argmin}_k \{|\mathbf{f}^L(\mathbf{p}^L(l)) - \mathbf{f}^L(k)|, \\ & k \in \mathcal{N}(\mathbf{p}^L(l)) \setminus \{\mathbf{p}^L(\nu),\ \nu = 1, \ldots, l\}\}.\end{aligned}$$

It can happen that the choice of the next index value $\mathbf{p}^L(l+1)$ is not unique, if the above minimum is attained for more than one index. In this case, one may fix favorite directions in order to determine a unique pathway.

Another situation which can occur during the procedure is that all indices in the neighborhood of an index $\mathbf{p}^L(l)$ have already been used in the path $\mathbf{p}^L$. In this case we have an interruption in the path vector. We need to choose one index $\mathbf{p}^L(l+1)$ from the remaining indices in J, which have not been taken yet in $\mathbf{p}^L$. There are different possibilities for finding a suitable next index. One simple choice is to take the smallest index from J that has not been used so far. Another choice is to look for a next index, such that again the absolute difference $|\mathbf{f}^L(\mathbf{p}^L(l)) - \mathbf{f}^L(\mathbf{p}^L(l+1))|$ is minimal, i.e., we take in this case

$$\begin{aligned}\mathbf{p}^L(l+1) \ = \ & \operatorname*{argmin}_k \{|\mathbf{f}^L(\mathbf{p}^L(l)) - \mathbf{f}^L(k)|, \\ & k \in J \setminus \{\mathbf{p}^L(\nu),\ \nu = 1, \ldots, l\}\}.\end{aligned}$$

By proceeding in this manner, we finally obtain a path vector $\mathbf{p}^L \in \mathbb{Z}^N$, which is a permutation of $(1, 2, \ldots, N)$.

After having constructed the path $\mathbf{p}^L$, we apply one level of the 1-D Haar DWT (or any other orthogonal or biorthogonal periodic DWT) to the vector of function values $(\mathbf{f}^L(\mathbf{p}^L(l)))_{l=1}^N$ along the path $\mathbf{p}^L$. We obtain the vector $\mathbf{f}^{L-1} \in \mathbb{R}^{N/2}$, containing the low-pass part, and the vector of wavelet coefficients $\mathbf{g}^{L-1} \in \mathbb{R}^{N/2}$. While the wavelet coefficients will be stored in $\mathbf{g}^{L-1}$, we further proceed with the low-pass vector $\mathbf{f}^{L-1}$ at the second level.

Further levels

If $N = 2^L r$ with $r \in \mathbb{N}$ being greater than or equal to the lengths of low-pass and high-pass filters in the chosen DWT, then we may apply the procedure L times. For a given vector $\mathbf{f}^{L-j}$, $0 < j < L$, at the $(j+1)$-th level we consider the index sets

$$J_l^{L-j} := J^{L-j+1}_{\mathbf{p}^{L-j+1}(2l-1)} \cup J^{L-j+1}_{\mathbf{p}^{L-j+1}(2l)},\ l = 1, \ldots, N/2^j,$$

with the corresponding function values $(\mathbf{f}^{L-j}(l))_{l=1}^{N/2^j}$. In particular, the index sets at the second level are $J_l^{L-1} := \{\mathbf{p}^L(2l-1), \mathbf{p}^L(2l)\}$, $l = 1, \ldots, N/2$, determining a partition of J.

We repeat the procedure described in the first step, but replacing the single indices with the new index sets J_l^{L-j}, and the corresponding function values with the smoothed function values $\mathbf{f}^{L-j}(l)$.

The new path vector $\mathbf{p}^{L-j} \in \mathbb{Z}^{N/2^j}$ should now be a permutation of $(1, 2, \ldots, N/2^j)$. We start again with the first index set J_1^{L-j}, i.e., $\mathbf{p}^{L-j}(1) = 1$. Having already found $\mathbf{p}^{L-j}(l)$, $1 \le l \le N/2^j - 1$, we determine the next value $\mathbf{p}^{L-j}(l+1)$ as

$$\mathbf{p}^{L-j}(l+1) = \operatorname*{argmin}_k \{|\mathbf{f}^{L-j}(\mathbf{p}^{L-j}(l)) - \mathbf{f}^{L-j}(k)|,$$
$$J_k^{L-j} \in \mathcal{N}(J^{L-j}_{\mathbf{p}^{L-j}(l)}) \setminus \{\mathbf{p}^{L-j}(\nu),\, \nu = 1, \ldots, l\}\}.$$

If the new value $\mathbf{p}^{L-j}(l+1)$ is not uniquely determined by the minimizing procedure, we can fix favorite directions in order to obtain a unique path. If for the set $J^{L-j}_{\mathbf{p}^{L-j}(l)}$ there is no neighboring index set that has not been used yet in the path vector $\mathbf{p}^{L-j}$, then we have to interrupt the path and to find a new good index set (that has been not used so far) to continue the path. As at the first level, we try to keep the differences of function values along the path as small as possible.

Finally, we apply the (periodic) wavelet transform to the vector $(\mathbf{f}^{L-j}(\mathbf{p}^{L-j}(l)))_{l=1}^{N/2^j}$ along the path $\mathbf{p}^{L-j}$, thereby obtaining the low-pass vector $\mathbf{f}^{L-j-1} \in \mathbb{R}^{N/2^{j+1}}$ and the vector of wavelet coefficients $\mathbf{g}^{L-j-1} \in \mathbb{R}^{N/2^{j+1}}$.

Output

As output of the complete procedure after L iterations we obtain the coefficient vector

$$\mathbf{g} = (\mathbf{f}^0, \mathbf{g}^0, \mathbf{g}^1, \ldots, \mathbf{g}^{L-1}) \in \mathbb{R}^N$$

and the vector determining the paths at each iteration step

$$\mathbf{p} = (\mathbf{p}^1, \mathbf{p}^2, \ldots, \mathbf{p}^L) \in \mathbb{R}^{2N(1-1/2^L)}.$$

These two vectors contain the entire information about the original function f.

In order to find a sparse representation of f, we apply a *shrinkage procedure* to the wavelet coefficients in the vectors $\mathbf{g}^j$, $j = 0, \ldots, L-1$ and obtain the vectors $\widetilde{\mathbf{g}}^j$.

Reconstruction

The reconstruction of $\mathbf{f}^L$ from $\widetilde{\mathbf{g}} = (\mathbf{f}^0, \widetilde{\mathbf{g}}^0, \widetilde{\mathbf{g}}^1, \ldots, \widetilde{\mathbf{g}}^{L-1})$ and $\mathbf{p}$ is given as follows.

$\widetilde{\mathbf{f}}^0 = \mathbf{f}^0$;

For $j = 0$ **to** $L-1$

- Apply the inverse DWT to the vector $(\widetilde{\mathbf{f}}^j, \widetilde{\mathbf{g}}^j) \in \mathbb{R}^{r2^j}$ in order to obtain $\widetilde{\mathbf{f}}_p^{j+1} \in \mathbb{R}^{r2^{j+1}}$.

- Apply the permutation $\widetilde{\mathbf{f}}^{j+1}(\mathbf{p}^{j+1}(k)) = \widetilde{\mathbf{f}}_p^{j+1}(k)$, for $k = 1, \ldots, r2^{j+1}$.

5. Example

We illustrate the simple idea of function decomposition with the EPWT on the sphere in the following small example. Let a set of 32 function values be given on the sphere, where each function value corresponds to a spherical triangle that has been obtained by radial projection of the triangulated octahedron in Figure 1 (left). The values are given as a vector $\mathbf{f} = \mathbf{f}^5$ of length 32, corresponding to the one-dimensional indexing of the triangles in Figure 1 (right),

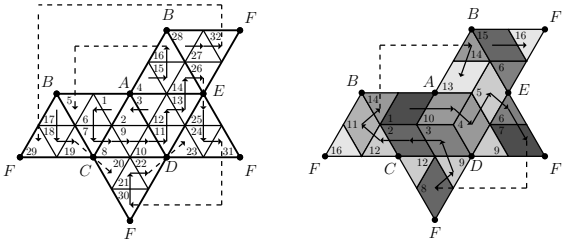


Figure 2. Illustration of first path through the triangulation $\mathcal{T}^1$ of the octahedron (left) and of the low-pass part after the first level of EPWT with Haar DWT (right). Index sets at the second level are illustrated by different gray values, and path vectors are represented by arrows.

$$\begin{aligned}\mathbf{f} = (&0.4492, 0.4219, 0.4258, 0.4375, 0.4141, 0.4531,\\ &0.4180, 0.4258, 0.4375, 0.4292, 0.4219, 0.4219,\\ &0.4219, 0.4258, 0.4023, 0.4141, 0.4219, 0.4219,\\ &0.4297, 0.4375, 0.4141, 0.4023, 0.4258, 0.4219,\\ &0.4258, 0.4180, 0.4531, 0.4141, 0.4375, 0.4258,\\ &0.4219, 0.4492).\end{aligned}$$

Starting with the index 1, with the function value 0.4492, we determine the first path vector. This index has the three neighbors $2, 4$, and 6, with the corresponding values 0.4219, 0.4375 and 0.4531, respectively (see Figure 2). Hence, the second index in the path is 6. Proceeding further according to Section 4 we obtain

$$\begin{aligned}\mathbf{p}^5 =&(1, 6, 7, 8, 9, 10, 11, 12, 13, 14, 26, 25, 24, 31, 30, 21,\\ &22, 23;\ 3, 2,\ 17, 18, 19, 20;\ 4, 15, 16, 5;\ 28, 27, 32, 29),\end{aligned}$$

where the interruptions in the path are indicated by semicolons. This path has four interruptions and is illustrated by arrows in Figure 2 (left). An application of the Haar DWT (with unnormalized filter coefficients $h_0 = h_1 = 1/2$, $g_0 = 1/2$, $g_1 = -1/2$) along this path gives (with truncation after four digits) the low-pass coefficients

$$\begin{aligned}\mathbf{f}^4 = (&0.4512, 0.4219, 0.4334, 0.4219, 0.4238, 0.4219,\\ &0.4219, 0.4200, 0.4140, 0.4238, 0.4219, 0.4336, 0.4199,\\ &0.4141, 0.4336, 0.4434),\end{aligned}$$

and the wavelet coefficients

$$\begin{aligned}\mathbf{g}^4 = (&-0.0020, -0.0039, -0.0042, 0., -0.0020,\\ &-0.0039, 0., 0.0058, -0.0118, 0.0020, 0., -0.0039,\\ &0.0176, 0., -0.0195, 0.0058).\end{aligned}$$

We now proceed to the second level. For the smoothed vector of function values $\mathbf{f}^4$ corresponding to the 16 index

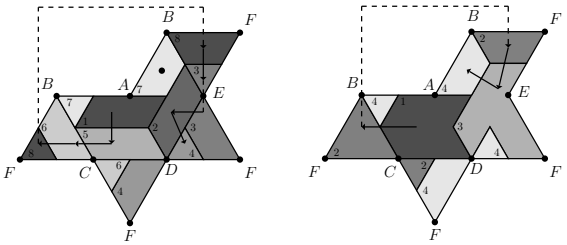


Figure 3. Illustration of the third and fourth paths.

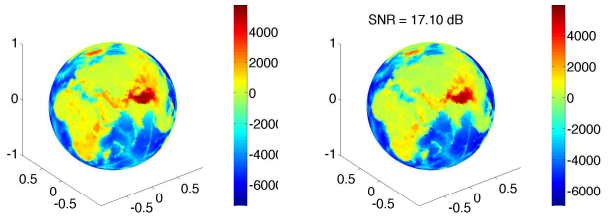


Figure 4. Approximation $\mathbf{f}^6$ at level 6 of the original dataset *topo* and the compressed version $\widetilde{\mathbf{f}}^6$ with threshold 2500.

sets that are illustrated by gray values in Figure 2 (right), we obtain the next path

$$\mathbf{p}^4 = (1, 10, 4, 5, 6, 7, 8, 9, 3, 2, 12, 11, 14, 13; 15, 16),$$

illustrated by arrows in Figure 2 (right). An application of the Haar DWT along $\mathbf{p}^4$ gives

$$\begin{aligned} \mathbf{f}^3 &= (0.4375, 0.4229, 0.4219, 0.4170, 0.4276, 0.4278, \\ &\quad 0.4170, 0.4385), \\ \mathbf{g}^3 &= (0.0136, -0.0010, 0., 0.0030, 0.0057, 0.0058, \\ &\quad 0.0029, -0.0049). \end{aligned}$$

At the third level we start with the smoothed vector $\mathbf{f}^3$ corresponding to the 8 index sets that are illustrated by gray values in Figure 3 (left). We find now the path $\mathbf{p}^3 = (1, 5, 6, 8, 3, 2, 4; 7)$, see Figure 3 (left). This leads to

$$\begin{aligned} \mathbf{f}^2 &= (0.4326, 0.4331, 0.4224, 0.4170), \\ \mathbf{g}^2 &= (0.0049, -0.0054, 0.0005, 0.). \end{aligned}$$

At the fourth level we have only 4 index sets that correspond to the values in $\mathbf{f}^2$, see Figure 3 (right). Hence we find $\mathbf{p}^2 = (1, 2, 3, 4)$ and

$$\mathbf{f}^1 = (0.4328, 0.4197), \qquad \mathbf{g}^1 = (-0.0003, 0.0027).$$

Finally, with $\mathbf{p}^1 = (1, 2)$, the last transform yields $\mathbf{f}^0 = (0.4263)$ and $\mathbf{g}^0 = (0.0066)$.

6. Numerical experiments

To illustrate the efficiency of our method, we took the dataset *topo* and we considered the regular octahedron with triangulation $\mathcal{T}_6$, containing 32768 triangles. The approximation $\mathbf{f}^6$ at level 6 is represented in Figure 4. We applied the EPWT with different thresholds, obtaining the compressed vector $\widetilde{\mathbf{f}}^6$, and we measured the SNR given as

$$SNR = 20 \cdot \log_{10} \frac{\|\mathbf{f}^6 - \mathrm{mean}(\mathbf{f}^6)\|_2}{\|\mathbf{f}^6 - \widetilde{\mathbf{f}}^6\|_2}.$$

threshold	number of remaining wavelet coeff.	l^2-norm of error	SNR
1	27732	26.4031	84.72
100	14185	5.34e+03	38.59
500	5230	2.47e+04	25.30
1000	3313	3.97e+04	21.17
1500	2699	5.00e+04	19.18
2000	2402	5.79e+04	17.89
2500	2265	6.35e+04	17.10

Table 1: Compression results for the dataset *topo*.

The results are contained in Table 1, where the mean of $\mathbf{f}^6$ is -2329.

Acknowledgments

This research in this paper is supported by the project 436 RUM 113/31/0-1 of the German Research Foundation (DFG). This is gratefully acknowledged.

References:

[1] R.L. Claypoole, G.M. Davis, W. Sweldens, and R.G. Baraniuk. Nonlinear wavelet transforms for image coding via lifting. *IEEE Trans. Image Process.* 12:1449–1459, 2003.

[2] A. Cohen and B Matei. Compact representation of images by edge adapted multiscale transforms. In *Proc. IEEE Int. Conf. on Image Process. (ICIP)*, Thessaloniki, pages 8–11, 2001.

[3] S. Dekel and D. Leviatan. Adaptive multivariate approximation using binary space partitions and geometric wavelets. *SIAM J. Numer. Anal.* 43:707–732, 2006.

[4] W. Ding, F. Wu, X, Wu, S. Li, and H. Li. Adaptive directional lifting-based wavelet transform for image coding. *IEEE Trans. Image Process.* 16:416–427, 2007.

[5] D.L. Donoho. Wedgelets: Nearly minimax estimation of edges. *Ann. Stat.* 27:859–897, 1999.

[6] S. Mallat. Geometrical grouplets. *Appl. Comput. Harmon. Anal.*, 26 (2): 143–290, 2009.

[7] G. Plonka. The easy path wavelet transform: A new adaptive wavelet transform for sparse representation of two-dimensional data. *Multiscale Model. Simul.* 7:1474–1496, 2009.

[8] D. Roşca. Haar wavelets on spherical triangulations. In Dodgson, N.A., Floater, M.S., Sabin, M.A., editors, *Advances in Multiresolution for Geometric Modelling*, Springer, pages 405–417, 2005.

[9] D. Roşca. Locally supported rational spline wavelets on a sphere. *Math. Comput.* 74:1803–1829, 2005.

[10] R. Shukla, P.L. Dragotti, M.N. Do, and M. Vetterli. Rate-distortion optimized tree structured compression algorithms for piecewise smooth images. *IEEE Trans. Image Process.* 14:343–359, 2005.

A fully non-uniform approach to FIR filtering

Brigitte Bidégaray-Fesquet [(1)] and Laurent Fesquet [(2)]

(1) LJK, CNRS / Grenoble University, B.P. 53, 38042 Grenoble Cedex 9, France.
(2) TIMA, 46 avenue Félix Viallet, 38031 Grenoble Cedex, France.
Brigitte.Bidegaray@imag.fr, Laurent.Fesquet@imag.fr

Abstract:

We propose a FIR filtering technique which takes advantage of the possibility of using a very low number of samples for both the signal and the filter transfer function thanks to non-uniform sampling. This approach leads to a summation formula which plays the role of the discrete convolution for usual FIR filters. Here the formula is much more complicated but it can be implemented and the evaluation of more elaborate expressions is compensated by the very low number of samples to process.

1. Introduction

Reducing the power consumption of mobile systems – such as cell phones, sensor networks and many others electronic devices – by one to two orders of magnitude is extremely challenging but will be very useful to increase the system autonomy and reduce the equipment size and weight. In order to reach such a goal, this paper proposes a solution applicable to FIR filtering which completely rethinks the signal processing theory and the associated system architectures.

Today the signal processing systems uniformly sample analog signals (at Nyquist rate) without taking advantage of their intrinsic properties. For instance, temperature, pressure, electro-cardiograms, speech signals significantly vary only during short moments. Thus the digitizing system part is highly constrained due to the Shannon theory, which fixes the sampling frequency at least twice the input signal frequency bandwidth. It has been proved in [4] and [6] that Analog-to-digital Converters (ADCs) using a non equi-repartition in time of samples leads to interesting power savings compared to Nyquist ADCs. A new class of ADCs called A-ADCs (for Asynchronous ADCs) based on level-crossing sampling (which produces non-uniform samples in time) [2, 3] and related signal processing techniques [1, 5] have been developed.

This work suggests an important change in the FIR filter design. As sampling analog signals is usually performed uniformly in time, sampling the filter transfer function is also done in a regular way with a constant frequency step. Non-uniform sampling leads to an important reduction of the weight-function coefficients. Combined with a non-uniform level-crossing sampling technique performed by an A-ADC, this approach drastically reduces the computation load by minimizing the number of samples and operations, even if they are more complex.

2. Principle and notations

For a large class of signal, non-uniform sampling leads to a reduced number of samples, compared to a Nyquist sampling. This feature has already been used in [1] to design non-uniform filtering techniques based on interpolation. In this work the authors however used a classical (uniform) filter, that is a usual discretization in time of the impulse response.

Here we want to go further and take advantage of the fact that the filter transfer function (the Fourier transform of the impulse response) is a very smooth function with respect to frequency. It can therefore be well approximated by the linear interpolation of quite few samples.

2.1 Level crossing sampling

The initial signals are supposed to be analog ones. The signal which we want to filter is given in the time domain and is denoted by $s(t)$. The filter transfer function is given in the frequency domain and is denoted by $H(\omega)$. The result of the filtering process $x(t)$ is then theoretically the convolution of $s(t)$ with the impulse response $h(t)$ which is the inverse Fourier transform of $H(\omega)$:

$$\begin{aligned} x(t) &= \int_{-\infty}^{+\infty} h(t-\tau)s(\tau)d\tau, \\ h(t) &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} H(\omega)e^{-i\omega t}d\omega. \end{aligned}$$

These signal are sampled in their initial domain using a level crossing scheme. This technique has to be adapted for the filter transfer function. Indeed level crossing has a sense if an order can be defined, for example for a real valued function. The filter transfer function is complex valued, therefore we can choose to sample either when the amplitude crosses some predefined values, or the phase, or both. The samples read $(s_n, \delta t_n)$ for the signal and $(H_k, \delta\omega_k)$ for the filter transfer function. These samples are formed of a value and the (time or frequency) interval length "elapsed" since the last sample. To give results or describe algorithms we will use the sample times or frequencies defined as $t_n = t_0 + \sum_1^n \delta t_{n'}$ and $\omega_k = \omega_0 + \sum_1^k \delta\omega_{k'}$ but computations will be performed using

only the time and frequency intervals δt_n and $\delta\omega_k$. We will also denote by $I_n = [t_{n-1}, t_n]$ and $J_k = [\omega_{k-1}, \omega_k]$ the time and frequency intervals.

2.2 Linear interpolation

To derive the FIR algorithm and approximate the theoretical integral formula, we form new analog functions from the previously described samples. To this aim we choose linear interpolation and we have

$$\begin{aligned}\bar{s}(t) &= \sum_n [a_n + b_n t]\chi_{I_n},\\ \bar{H}(\omega) &= \sum_k (\alpha_k + \beta_k\omega)e^{i(\gamma_k+\delta_k\omega)}\chi_{J_k},\end{aligned}$$

where χ denotes the indicator function of the set given in index. The coefficients a_n and b_n can be expressed in terms of s_n, s_{n-1}, t_n and δt_n. The coefficients α_k, β_k, γ_k and δ_k can be expressed in terms of H_k, H_{k-1}, ω_k and $\delta\omega_k$.

In fact these formulae cover the piecewise constant case (only take $b_n = \beta_k = \delta_k = 0$) in three possible forms: constant on intervals I_n or nearest neighbor interpolation, with a possible need to modify the definition of t_n and δt_n in the algorithms. They also cover two ways to linearly interpolate the complex valued filter transfer function: either interpolate separately the amplitude and the phase (α_k and β_k are real) or interpolate in the complex plane (α_k and β_k are complex, γ_k and δ_k are zero).

The digital filter then consists in computing (possibly) for all time

$$\begin{aligned}\bar{x}(t) &= \int_{-\infty}^{+\infty} \bar{h}(t-\tau)\bar{s}(\tau)d\tau,\\ \bar{h}(t) &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} \bar{H}(\omega)e^{-i\omega t}d\omega.\end{aligned}$$

3. Deriving a filtering formula in the general context

3.1 A summation formula

The impulse response $\bar{h}(t)$ can be split in contributions for each frequency sample $\bar{h}(t) = \sum_k h_k(t)$ with

$$h_k(t) = \frac{1}{2\pi}\int_{\omega_{k-1}}^{\omega_k} (\alpha_k + \beta_k\omega)e^{i(\gamma_k+\delta_k\omega)}e^{-i\omega t}d\omega$$

for which we will give an explicit expression in Section 3.2. Although the piecewise linear function $\bar{H}(\omega)$ has a compact support (we only have a finite number of samples), the functions $h_k(t)$ have an infinite support. This is not a problem since the convolution will involve $\bar{s}(t)$ which has a compact support. The convolution reads

$$\begin{aligned}\bar{x}(t) &= \int_{-\infty}^{+\infty} \bar{h}(t-\tau)\bar{s}(\tau)d\tau\\ &= \sum_n \int_{t_{n-1}}^{t_n} h(t-\tau)s_n(\tau)d\tau\\ &= \sum_n\sum_k \int_{t_{n-1}}^{t_n} h_k(t-\tau)(a_n + b_n\tau)d\tau\\ &= \sum_n \left(a_n \sum_k h^0_{nk}(t) + b_n \sum_k h^1_{nk}(t)\right)\end{aligned}$$

where

$$h^0_{nk}(t) = \int_{t_{n-1}}^{t_n} h_k(t-\tau)d\tau,$$

$$h^1_{nk}(t) = \int_{t_{n-1}}^{t_n} h_k(t-\tau)\tau d\tau.$$

We obtain a summation formula as in the classical FIR filtering case where it takes the form of a discrete convolution. To be closer to this classical case, we should write this as

$$\bar{x}(t) = \sum_n s_n \sum_k h_{nk}(t),$$

which is possible but the effective expression depends on the type of interpolation used (piecewise constant or linear).

There remains to make explicit these two types of elementary contributions.

3.2 Elementary impulse responses

A straightforward computation of the integral formulation for $h_k(t)$ yields

$$\begin{aligned}h_k(t) &= \frac{\alpha_k e^{i\gamma_k}}{2\pi}\int_{\omega_{k-1}}^{\omega_k} e^{i(\delta_k-t)\omega}d\omega\\ &+ \frac{\beta_k e^{i\gamma_k}}{2\pi}\int_{\omega_{k-1}}^{\omega_k} e^{i(\delta_k-t)\omega}\omega d\omega\\ &= \frac{\alpha_k e^{i\gamma_k}\left(e^{i(\delta_k-t)\omega_k} - e^{i(\delta_k-t)\omega_{k-1}}\right)}{2\pi i(\delta_k-t)}\\ &+ \frac{\beta_k e^{i\gamma_k}\left(\omega_k e^{i(\delta_k-t)\omega_k} - \omega_{k-1}e^{i(\delta_k-t)\omega_{k-1}}\right)}{2\pi i(\delta_k-t)}\\ &+ \frac{\beta_k e^{i\gamma_k}\left(e^{i(\delta_k-t)\omega_k} - e^{i(\delta_k-t)\omega_{k-1}}\right)}{2\pi(\delta_k-t)^2}.\end{aligned}$$

These formulae seem singular when $t = \delta_k$. This is not the case and has no reason to be since the function we integrate is smooth with respect to all parameters and variables. The limiting value for $t = \delta_k$ is clearly

$$\begin{aligned}h_k(\delta_k) &= \frac{\alpha_k e^{i\gamma_k}}{2\pi}\int_{\omega_{k-1}}^{\omega_k} d\omega + \frac{\beta_k e^{i\gamma_k}}{2\pi}\int_{\omega_{k-1}}^{\omega_k}\omega d\omega\\ &= \frac{e^{i\gamma_k}}{2\pi}\delta\omega_k(\alpha_k + \beta_k\frac{1}{2}(\omega_{k-1}+\omega_k)).\end{aligned}$$

3.3 Elementary summation coefficients

A quick glance at the explicit expression of $h_k(t)$ clearly provides the impression that the explicit formulae for $h^0_{nk}(t)$ and $h^0_{nk}(t)$ will not fit in the columns here. We will give only their flavor. Indeed we want to compute the time integrals of of $h_k(t-\tau)$ and $h_k(t-\tau)\tau$ for $\tau \in I_n$. This leads to integrate the product of a rational function with a complex exponential function. The results cannot be given in terms of simple functions but only in terms of the exponential integral function

$$\mathrm{Ei}(ix) = -\int_x^\infty e^{iy}\frac{dy}{y} + i\frac{\pi}{2}.$$

We give in the next section a simple example of elementary summation coefficient calculation in the piecewise linear context.

4. A simple and ideal example

4.1 Computation of the coefficients

Our sampling for the filter transfer function yields a particularly simple formulation for the ideal low-pass filter which is 1 on the frequency interval $[-\omega_c, \omega_c]$ and zero elsewhere. This yields a single sample $(1, 2\omega_c)$ and linearly interpolated coefficients $\alpha_1 = 1$, $\beta_1 = 0$, $\gamma_1 = 0$ and $\delta_1 = 0$. The expression for the elementary impulse response is

$$h_1(t) = \frac{\left(e^{-i\omega_c t} - e^{i\omega_c t}\right)}{-2\pi i t} = \frac{\omega_c}{\pi}\,\mathrm{sinc}(\omega_c t).$$

Then we have to compute

$$\begin{aligned} h^0_{n1}(t) &= \int_{t_{n-1}}^{t_n} h_1(t-\tau)d\tau = -\int_{t-t_{n-1}}^{t-t_n} h_1(\tau)d\tau \\ &= -\frac{1}{\pi}(\mathrm{Si}(\omega_c(t-t_n)) - \mathrm{Si}(\omega_c(t-t_{n-1})), \end{aligned}$$

where Si is the special function known as sine integral and defined by

$$\mathrm{Si}(x) = \int_0^x \sin(y)\frac{dy}{y} = \frac{1}{2i}(\mathrm{Ei}(ix) - \mathrm{Ei}(-ix)) + \frac{\pi}{2},$$

and

$$\begin{aligned} h^1_{n1}(t) &= \int_{t_{n-1}}^{t_n} h_1(t-\tau)\tau d\tau \\ &= -\int_{t-t_{n-1}}^{t-t_n} h_1(\tau)(t-\tau)d\tau \\ &= t\,h^0_{n1}(t) + \frac{1}{\pi}\int_{t-t_n}^{t-t_{n-1}} \sin(\omega_c\tau)d\tau \\ &= t\,h^0_{n1}(t) - \frac{1}{\pi\omega_c}(\cos(\omega_c(t-t_n)) \\ &\qquad - \cos(\omega_c(t-t_{n-1}))). \end{aligned}$$

This case is simple due to its minimal number of samples in the frequency domain, but it displays all the difficulties of the general case, i.e. the need to evaluate special functions. These functions are built in many libraries in view of a numerical implementation of these algorithms. Moreover these functions are however very smooth: the Si function for example is almost linear in the neighborhood of 0 and tends to $\pm\pi/2$ at $\pm\infty$ with very gentle oscillations. This feature makes possible the construction of efficient lookup tables in view of a hardware implementation.

4.2 Numerical results

To illustrate this simple example we filter the signal

$$s(t) = 0.45\sin(2\pi t) + 0.45\sin(10\pi t) + 0.9$$

with the ideal low pass filter with the cutoff frequency $\omega_c = 4\pi$. The theoretical result is therefore supposed to be

$$x(t) = 0.45\sin(2\pi t) + 0.9.$$

This is not the typical sort of signal which is supposed to be addressed by our technique since it is not a sporadic one and a relatively large number of samples are taken. We perform the computations within the MATLAB SPASS (Signal Processing for ASynchronous Systems) framework (`http://ljk.imag.fr/membres/Brigitte.Bidegaray/SPASS/`).
This signal is sampled with a M-bit Asynchronous A/D Converter (AADC) which leads to a level crossing sampling over the amplitude range $[0, 1.8]$.
We can choose as we want the times at which the filtered signal is computed. To display the results we choose the sequence of times $t_m = .17m$ (m integer) to have sampling points dispatched irregularly over the obtained solution.
On Figure 1, you can see the result for a linear interpolation of the signal non-uniform samples and a 3-bit AADC. We plot continuous functions with lines: the initial signal $s(t)$ (dashed line) and the theoretical filtered signal $x(t)$ (solid line). We plot the sampled results with markers: the non-uniformly sampled initial signal s_n (asterisk markers) and the computed filtered samples x_m (circle markers) at times t_m.

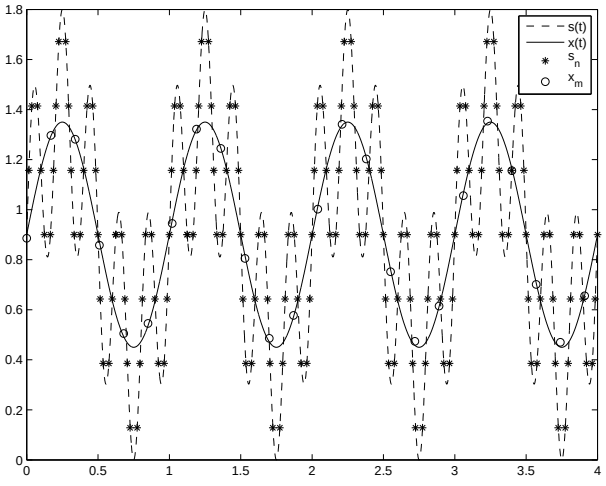


Figure 1: Filtering result. Initial signal (dashed line), theoretical filtered signal (solid line), non-uniformly sampled initial signal (asterisk markers) and computed filtered samples (circle markers).

This very simple test case has quite a low number of parameters compared to the full problem for which we can finely tune the filter transfer function sampling for example. We compare here the results obtained for a zeroth and a first order interpolation of the signal and for different values (2, 3, 4 and 5) of the AADC resolution. On Table 1 we give the relative l^1 error between computed filtered samples x_m at times $t_m = .01m$ (m integer) and the theoretical values $x(t_m)$.

	0th order	1st order
$M = 2$	0.0608	0.0584
$M = 3$	0.0076	0.0046
$M = 4$	0.0052	0.0045
$M = 5$	0.0046	0.0045

Table 1: l^1 error of the filtering method for 0th and first order interpolation of the signal and and M bit resolution of the AADC ($M = 2,\ 3,\ 4,\ 5$).

In the case of the 2-bit AADC, there are 2.8 points per wavelength for the highest frequency part of the signal. This is a very low rate, and we are however able to have only 6% error on the filtered result which is quite sufficient for a large range of applications. The other results all show less than 1% error. The values displayed on Table 1 are very dependent on the choice of the function to filter. Finer results (allowing less than .45% error) should certainly be obtained by using a higher order interpolation for the signal.

5. Conclusions

We have presented a novel approach to FIR filtering based on the non-uniform sampling of the signal but also the non-uniform sampling in frequency of the filter transfer function. The final result is complex but is nonetheless possible to implement in hardware devices and of course in numerical codes. This complexity is balanced by the very low number of samples and the relatively low number of operations needed for each evaluation. This approach is very promising to achieve a lower power consumption in mobile systems.

6. Acknowledgments

This work has been supported by a funding from the Joseph Fourier-Grenoble 1 University: MSTIC project TATIE.

References:

[1] Fabien Aeschlimann, Emmanuel Allier, Laurent Fesquet, and Marc Renaudin. Asynchronus fir filters, towards a new digital processing chain. In *10th IEEE International Symposium on Asynchronous Circuits and Systems (ASYNC'04)*, pages 198–206, Crete, Greece, April 2004.

[2] Filipp Akopyan, Rajit Manohar, and Alyssa B. Apsel. A level-crossing flash asynchronous analog-to-digital converter. In *12th IEEE International Symposium on Asynchronous Circuits and Systems (ASYNC'06)*, pages 11–22, Grenoble, France, March 2006.

[3] Emmanuel Allier, Gilles Sicard, Laurent Fesquet, and Marc Renaudin. A new class of asynchronous A/D converters based on time quantization. In *9th IEEE International Symposium on Asynchronous Circuits and Systems (ASYNC'03)*, pages 197–205, Vancouver, Canada, May 2003.

[4] Jon W. Mark and Terence D. Todd. A nonuniform sampling approach to data compression. *IEEE Trans. on Communications*, COM-29(1):24–32, January 1981.

[5] Saeed Mian Qaisar, Laurent Fesquet, and Marc Renaudin. Adaptive rate filtering for a signal driven sampling scheme. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP07)*, pages III–1465–III–1468, Honolulu, Hawaii, USA, April 2007.

[6] N. Sayiner, H.V. Sorensen, and T.R. Viswanathan. A level-crossing sampling scheme for A/D conversion. *IEEE Trans. on Circuits and Systems II*, 43(4):335–339, April 1996.

ADAPTIVE TRANSMISSION FOR LOSSLESS IMAGE RECONSTRUCTION

Elisabeth Lahalle, Gilles Fleury, Rawad Zgheib

Department of Signal Processing and Electronic Systems, Supélec, Gif-sur-Yvette, France
E-mail : firstname.lastname@supelec.fr
tel: +33 (0)1 69 85 14 27, fax: +33 (0)1 69 85 14 29

ABSTRACT

This paper deals with the problem of adaptive digital transmission systems for lossless reconstruction. A new system, based on the principle of non-uniform transmission, is proposed. It uses a recently proposed algorithm for adaptive stable identification and robust reconstruction of AR processes subject to missing data. This algorithm offers at the same time an unbiased estimation of the model's parameters and an optimal reconstruction in the least mean square sense. It is an extension of the RLSL algorithm to the case of missing observations combined with a Kalman filter for the prediction. This algorithm has been extended to 2D signals. The proposed method has been applied for lossless image compression. It has shown an improvement in bit rate transmission compared to the JPEG2000 as well as the JPEG-LS standards.

Index Terms— adaptive, lossless, compression

1. INTRODUCTION

Lossless compression methods are important in many medical applications where large data set need to be transmitted without any loss of information. Actually, some lesions risk becoming undetectable due to the effects of lossy compression. General lossless compression coders are considered to be composed of two main blocks: a data decorrelation block and an entropy coder for the decorrelated data. Two main tendencies may be noticed for the methods used for the decorrelation step: methods based on wavelet transforms and methods based on predictive coding. They have led to the main compression standards : the JPEG2000 for the former group of methods [1], the JPEG-LS for the latter [2]. Intensive attention is paid to transform based compression methods with many algorithms which perform well regarding the bit rate such as SPIHT [3], QT [4], etc.
All these coders use a uniform transmission of the binary elements to transmit. In a previous paper [5], the design of digital systems based upon non-uniform transmission of signal samples was introduced. The idea behind is to avoid sending a sample if it can be efficiently predicted, e.g. with a prediction error smaller than the quantization one, thus reducing the average transmission bit rate and increasing the signal to noise ratio (SNR). A speech coder based on the Adaptive Pulse Code Modulation (ADPCM) principle and non-uniform transmission of signals have already been proposed in [6]. It uses the Least Mean Square (LMS)-like algorithm [7] for the prediction of the samples that were not sent. However, this algorithm converges toward biased estimations of the model's parameters and does not use an optimal predictor in the least mean square sense. Recently, we proposed a Recursive Least Square Lattice (RLSL) algorithm for adaptive stable identification of non stationary Autoregressive (AR) processes subject to missing data, using a Kalman filter as a predictor [8]. This algorithm is fast, guarantees the stability of the model identified and offers at the same time an optimal reconstruction error in the least mean square sense and an unbiased estimation of the model's parameters in addition to the fast adaptivity to the variations of the parameters in the case of non stationary processes. Non stationnary AR processes can model a large number of signals in practical situations, such as images in the bi-dimensional case [9]. A new lossless image coder based on a non-uniform transmission principle is proposed: it is based on an adaptation of the algorithm proposed in [8] for optimal prediction and identification of 2D AR processes subject to missing observations.

In the following, begin by presenting the non-uniform transmission idea for lossless compression. In a second part, the adaptive algorithm for reconstruction of AR processes with missing observations [8] is described and extended to 2D AR processes. Its integration into a non-uniform transmission system is studied in the third section. Finally, an example illustrates the performances of the proposed system. It is compared to a uniform digital transmission system : the JPEG2000.

2. NON-UNIFORM TRANSMISSION SYSTEM FOR LOSSLESS RECONSTRUCTION

The proposed system uses predictive coding and non-uniform transmission to reduce the bit rate transmission. An AR signal modeling is considered for the prediction. Let x_n be the amplitude of the signal at time n. The prediction of a sample will be noted $\hat{x}_{n,P}$ and the prediction error $e_{n,P} = x_n - \hat{x}_{n,P}$. In the receiver, a sample x_n is predicted using the estimated model parameters at time $n-1$, $\hat{\mathbf{a}}_{n-1}$, and the available sam-

ples. The key ideas of the proposed system are the following. If $e_{n,P} \approx 0$, x_n is replaced by $\hat{x}_{n,P}$ in the receiver without any loss, requiring only one bit flag to be transmitted for the first and the last sample where $e_{n,P} \approx 0$. If an efficient prediction method for non-uniformly sampled data is used, the above situation occurs many times during the transmission. This is the case for example outside the region of interest of the image where the sample value is constant or null. The whole number of transmitted samples is thus considerably reduced. As some of the samples are not transmitted, the receiver has to deal with the problem of online identification and reconstruction of signals subject to missing samples. The probability law of the prediction error of the image to transmit is then used to adapt the number of bit coding the prediction error in the case where it is non zero.

3. PREDICTION/RECONSTRUCTION FOR NON-UNIFORMLY SAMPLED DATA

3.1. Kalman RLSL algorithm

Let $\{x_n\}$ be an AR process of order L with parameters $\{a_k\}$, and $\{\epsilon_n\}$ the corresponding innovation process of variance σ_ϵ^2. The loss process is modeled by an i.i.d binary random variable $\{c_n\}$, where $c_n = 1$ if x_n is available, otherwise $c_n = 0$. Let $\{z_n\}$ be the reconstruction of the process $\{x_n\}$. If x_n is available $z_n = x_n$, otherwise, $z_n = \hat{x}_n$, the prediction of x_n. In order to identify, in real time, the AR process subject to missing data, the algorithm proposed in [8] can be summarised as follows. The reflection coefficients of the lattice structure are determined by minimizing the weighted sum of the quadratic forward, $f_t^{(l)}$, and backward, $b_t^{(l)}$, prediction errors :

$$E_n^{(l)} = \sum_{i=1}^{n} w_{n-i} \left(f_n^{(l)2} + b_n^{(l)2} \right). \tag{1}$$

A Kalman filter provide an optimal prediction of the signal using the AR estimated parameters. These parameters are deduced from the estimated reflection coefficients using the Durbin Levinson recursions. At time $n+1$, the first line of the matrix A of the state space representation of an AR process is built with $\hat{\mathbf{a}}_n^{(L)\top}$, the vector of the parameters estimated at time n. The matrix is then named A_{n+1}.

$$A_{n+1} = \left[\begin{array}{ccc|c} \hat{a}_{1,n}^{(L)} & \dots & \dots & \hat{a}_{L,n}^{(L)} \\ \hline 1 & & 0 & 0 \\ & \ddots & & \vdots \\ 0 & & 1 & 0 \end{array}\right], \tag{2}$$

$$P_{n+1|n} = A_{n+1} P_{n|n} A_{n+1}^\top + R_\epsilon,$$
$$\hat{\mathbf{x}}_{n+1|n} = A_{n+1} \hat{\mathbf{x}}_{n|n}$$
$$\hat{y}_{n+1|n} = c_{n+1} \hat{x}_{n+1|n}$$

If x_{n+1} is available, i.e. $c_{n+1} = 1$,

$$K_{n+1} = P_{n+1|n} \mathbf{c}_{n+1} (\mathbf{c}_{n+1}^\top P_{n+1|n} \mathbf{c}_{n+1})^{-1}, \tag{3a}$$
$$P_{n+1|n+1} = (I_d - K_{n+1} \mathbf{c}_{n+1}^\top) P_{n+1|n}, \tag{3b}$$
$$\hat{\mathbf{x}}_{n+1|n+1} = \hat{\mathbf{x}}_{n+1|n} + K_{n+1} (y_{n+1} - \hat{y}_{n+1|n}) \tag{3c}$$

The predictions of the previous missing data up to time $n - L + 1$ are updated thanks to the filtering of the state in equation (3c). It is convenient now to calculate all the variables of the lattice filter since the last available observation at time $n - h$, where $h \geq 0$ depends on the observation pattern. At each time t, for $n - h + 1 \leq t \leq n + 1$, the recursive equations of the RLSL algorithm given by (5) are applied to estimate the different reflection coefficients $\hat{k}_t^{(l)}$ and prediction errors $\hat{f}_t^{(l)}, \hat{b}_t^{(l)}$ for $1 \leq l \leq L$. The values of the forward and backward prediction errors are initialized using the updated estimates of the missing samples (those contained within the filtered state $\hat{\mathbf{x}}_{n+1|n+1}$), i.e. $\hat{f}_t^{(0)} = \hat{b}_t^{(0)} = \hat{x}_{t|n+1}$.

Hence,

- For $t = n - h + 1$ to $n + 1$
  - Initialize for $l = 0$
$$\hat{f}_t^{(0)} = \hat{b}_t^{(0)} = \hat{x}_{t|n+1}, \hat{k}_t^{(0)} = 1, \tag{4}$$
  - For $l = 1$ to $\min(L, n)$
$$C_t^{(l)} = \lambda C_{t-1}^{(l)} + 2 \hat{f}^{(l-1)_t} \hat{b}_{t-1}^{(l-1)}, \tag{5a}$$
$$D_t^{(l)} = \lambda D_{t-1}^{(l)} + \hat{f}_t^{(l-1)2} + \hat{b}_{t-1}^{(l-1)2}, \tag{5b}$$
$$\hat{k}_t^{(l)} = -\frac{C_t^{(l)}}{D_t^{(l)}}, \tag{5c}$$
$$\hat{f}_t^{(l)} = \hat{f}_t^{(l-1)} - \hat{k}_t^{(l)} \hat{b}_{t-1}^{(l-1)}, \tag{5d}$$
$$\hat{b}_t^{(l)} = \hat{b}_{t-1}^{(l-1)} - \hat{k}_t^{(l)} \hat{f}_t^{(l-1)}, \tag{5e}$$
  - end
- end.

The AR parameters at time $n+1$, $(\hat{a}_{i,n+1}^{(L)})_{1 \leq i \leq L}$, are deduced from the reflection coefficients $(\hat{k}_{n+1}^{(l)})_{1 \leq i \leq L}$ using the Durbin Levinson recursions. owever if x_{n+1} is absent, $c_{n+1} = 0$, the predicted state, $\hat{\mathbf{x}}_{n+1|n}$, is not filtered by the Kalman filter, and the parameters are not updated since the reflection coefficients $(\hat{k}_{n+1}^{(l)})_{1 \leq l \leq L}$ are not yet calculated,

$$K_{n+1} = 0, \tag{6a}$$
$$P_{n+1|n+1} = P_{n+1|n}, \tag{6b}$$
$$\hat{\mathbf{x}}_{n+1|n+1} = \hat{\mathbf{x}}_{n+1|n}, \tag{6c}$$
$$\hat{\mathbf{a}}_{n+1}^{(L)} = \hat{\mathbf{a}}_n^{(L)}. \tag{6d}$$

The cost function minimized by this algorithm is the weighted mean of all quadratic prediction errors. When a sample is

missing, the prediction error can not be calculated, it is replaced by its estimation. Indeed, recall that in order to update the reflection coefficients at a time n, the lattice filter variables must have been calculated at all previous times. Therefore, using this algorithm, the lattice filter variables are estimated at all times even when a sample is missing. Consequently, this algorithm presents an excellent convergence behavior and have fast parameter tracking capability even for a large probability of missing a sample. The computational complexity of this algorithm is found to be $O((1-q)NL^2)$, where q is the bernoulli's probability of losing a sample, N is the size of the signal and L the order of the AR model.

3.2. Adaptation to 2D signals

A first solution to use the previous algorithm for 2D signals is to use the classical video scanning of the image in order to get a 1D signal. However, only a 1D decorrelation is achieved using this method.
In order to get a 2D decorrelation of the image, a 2D AR predictor $\hat{x}_{i,j}$ of the sample $x_{i,j}$ (7) must be used in addition to the video scanning of the image.

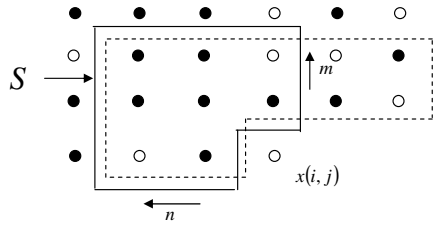


Fig. 1. AR 2D: prediction support

$$\hat{x}_{i,j} = \sum_{n,m \in S} \hat{a}_{n,m} x_{i-n,j-m} \tag{7}$$

In order to integrate this 2D AD predictor into the previous algorithm, the first line of the A matrix is built with the $\hat{a}_{n,m}$ parameters, and the regressor vector $[x_{n-1} \dots x_{n-L}]^\top$ is replaced by $[x_{i-1,j} \dots x_{i-n,j-m} \dots x_{i-p,j-q}]^\top$. The renumbering task excepted, to built the A matrix, the computational time of these 2D algorithm is similar to the 1D one.

4. PROPOSED ADAPTATIVE TRANSMISSION ALGORITHM

In this section, we propose to use the algorithms discussed in section 3 as efficient predictors in the non uniform transmission system proposed in section 2 in order to minimize the number of bit to transmit. At each time n, knowing all transmitted samples and using the same identification and reconstruction method as the one used in the receiver, the transmitter evaluates the signal reconstruction performance in the receiver. This can be done by comparing the receiver prediction error, $|e_{n,P}|$, with different thresholds, $\mathcal{S}_1 \approx 0, \mathcal{S}_2, ..., \mathcal{S}_i$. Thus, if the receiver is able to reconstruct the sample without error (error greatly smaller than the quantification error ($1e^{-5}$)), only a one bit flag is transmitted to indicate the first and the last missing sample. The number of thresholds $\mathcal{S}_i$ and their values are chosen according to the probability law of the prediction error to transmit only the B_i bits required to code the prediction error for each threshold. The proposed coding decoding algorithm can be summarized, at a time n, as:

- In the transmitter:
  . $e_{n,P} = x_n - \hat{x}_{n,P}$
  . if ($|e_{n,P}| = 1e^{-5}$ and $|e_{n-1,P}| > 1e^{-5}$ or $|e_{n,P}| > 1e^{-5}$ and $|e_{n-1,P}| = 1e^{-5}$), one bit flag is transmitted,
  . else if $|e_{n,P}| < \mathcal{S}_2$,
  . if $|e_{n,P}| < \mathcal{S}_3$, B_3 bits are transmitted,
  . else B_2 bits are transmitted,
  . else B_1 bits are transmitted.

- In the receiver, the method described in 3 is used for adaptive identification and reconstruction of a signal subject to missing data: if a new sample is received, the AR parameters are updated. Otherwise, the missing sample is predicted in terms of the past available samples and the current estimation of the parameters.

5. SIMULATIONS

The performances of both proposed methods are compared to the JPEG2000. The first method uses a 1D AR model of order 3 of the signal. In the second method, the image is modeled by a 2D AR process of order $(2, 2)$. The performances of the different methods are evaluated in term of bit rate (in bpp) on CT images. The PSNR is computed for the proposed methods to show the lossless reconstruction of the image. The PSNR which have been reached for all the simulations corresponds to the infinity value. Table 1 shows the results for CT images of (512x512x12) bits presented in figures 2, 3 and 4 (Images courtesy of Dr Kopans, MGH Boston, USA. Tomosynthesis investigational device from GE Healthcare (Chalfont St Giles, UK)). In these images the prediction error is in most of the case small (lower than 32), but for the pixels of the edge of the ROI the prediction error requires 12 bits to be coded. Consequently, the following values are chosen for the number of bit to code the prediction error : $B_1 = 13$, $B_2 = 8$, $B_3 = 6$.

6. CONCLUSION

A new digital transmission system for lossless image reconstruction has been proposed. It is based on a non-uniform transmission principle and on extensions to 2D of the algorithm proposed in [8] for real time identification and reconstruction of AR processes subject to missing data. The pro-

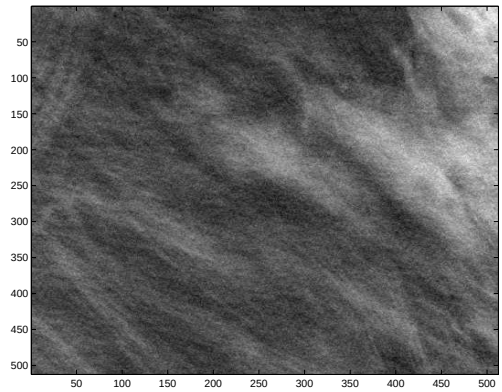


Fig. 2. CT1 image

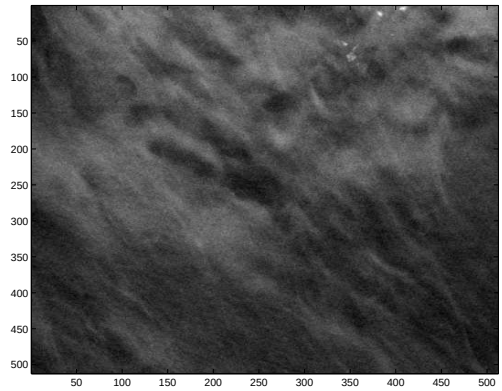


Fig. 3. CT2 image

posed methods, applied on CT images, has shown in their two forms (2D as well as 1D) an improvement in bit rate comparing to the JPEG2000 and JGPEG-LS standards. Comparing to the JPEG2000, significant gains for lossless compression are reached: 3.4% for CT3 image up to 4.6% for CT1 image. Comparing to the JPEG-LS, the most significant gains (2.7% up to 3.6%) are reached for CT2 and CT1 images where the RLE coding of the JPEG-LS is not used.

7. REFERENCES

[1] ISO/IEC 15444-1, "Information technology - jpeg2000 image coding system," JPEG2000 standard, Part 1-Core coding system, 2000.

Table 1. Comparison of the three methods in bit rate (in bpp) for CT images of (512x512x12) bits:

Method	$CT1$	$CT2$	$CT3$
1	6.53	6.67	5.10
2	6.45	6.61	5.10
JPEG2000	6.76	6.89	5.28
JPEG-LS	6.69	6.79	5.15

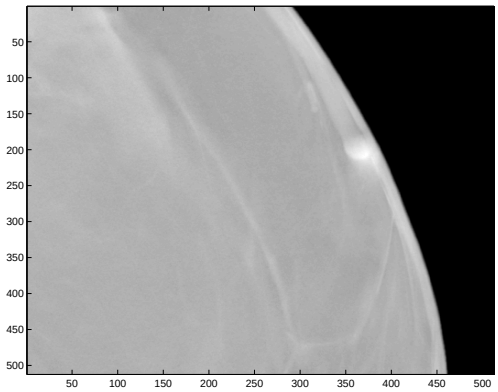


Fig. 4. CT3 image

[2] ISO/IEC 14495-1, "Information technology - lossless and near-lossless compression of continuous-tone still images," JPEG-LS standard, Baseline, 2000.

[3] A. Said and W. A. Pearlman, "A new fast and efficient image codec based on set partitionning in hierarchical trees," *IEEE Trans. on Circuits and systems for Video Technology*, vol. 6, pp. 243–250, June 1996.

[4] A. Munteanu and J. Cornelis, "Wavelet based lossless compression scheme with progressive transmission capability," *International Journal of Imaging Systems and Tecnology*, vol. 10, pp. 76–85, January 1999.

[5] S. Mirsaidi, G. Fleury, and J. Oksman, "Reducing quantization error using prediction/non uniform transmission," in *Proc. International Workshop on Sampling Theory and Applications*. IEEE, 1997, pp. 139–143.

[6] E. Lahalle and J. Oksman, "ADPCM speech coder with adaptive transmission and ARMA modelling of non-uniformly sampled signals," in *5th Nordic Signal Processing Symposium, CD-ROM proceedings, Norway*. IEEE, 2002.

[7] S. Mirsaidi, G. Fleury, and J. Oksman, "LMS like AR modeling in the case of missing observations," *IEEE Transactions on Signal Processing*, vol. 45, pp. 1574–1583, June 1997.

[8] R. Zgheib, G. Fleury, and E. Lahalle, "Lattice algorithm for adaptive stable identification and robust reconstruction of non stationary ar processes with missing observations," *IEEE Transactions on Signal Processing*, vol. 56, pp. 2746–2754, July 2008.

[9] N. S. Jayant and P. Noll, "Digital coding of waveform, principles and applications to speech and video," Prentice Hall, 1984.

Geometric Sampling of Images, Vector Quantization and Zador's Theorem

Emil Saucan [(1)], Eli Appleboim [(2)] and Yehoshua Y. Zeevi [(2)]

(1) Department of Mathematics, Technion - Israel Institute of Technology, Haifa 32000, Israel.
(3) Electrical Engineering Department, Technion - Israel Institute of Technology, Haifa 32000, Israel.
semil@tx.technion.ac.il, eliap@ee.technion.ac.il, zeevi@ee.technion.ac.il

Abstract:

We present several consequences of the geometric approach to image sampling and reconstruction we have previously introduced. We single out the relevance of the geometric method to the vector quantization of images and, more important, we give a concrete and candidate for the optimal embedding dimension in Zador's Thorem. An additional advantage of our approach is that that this provides a constructive proof of the aforementioned theorem, at least in the case of images. Further applications are also briefly discussed.

1. Introduction

In recent years it became common amongst the signal processing community, to consider images and other signals as well, as Riemannian manifolds embedded in higher dimensional spaces. Usually, the embedding manifold is taken to be $\mathbb{R}^n$, but other options can, and had been considered. Along with that, sampling is an essential preliminary step in processing of any continuous signal by a digital computer. This step lies at heart of any digital processing of any (presumably continuous) data/signal. It is therefore natural to strive to achieve a sampling method for images, viewed as such, that is as higher dimensional dimensional objects (i.e. manifolds), rather than their representation as 1-dimensional signals. In consequence, our sampling and reconstruction techniques stem from the the fields of differential geometry and topology, rather than being motivated by the traditional framework of harmonic analysis. More precisely, our approach to Shannon's Sampling Theorem is based on sampling the graph of the signal, considered as a manifold, rather than a sampling of the domain of the signal, as is customary in both theoretical and applied signal and image processing. In this context it is important to note that Shannon's original intuition was deeply rooted in the geometric approach, as exposed in his seminal work [14].
Our approach is based upon the following sampling theorem for differentiable manifolds that was recently presented and applied in the context image processing [12]:

Theorem 1 *Let $\Sigma^n \subset \mathbb{R}^N, n \geq 2$ be a connected, not necessarily compact, smooth manifold, with finitely many compact boundary components. Then, there exists a sampling scheme of Σ^n, with a metric density $\mathcal{D} = \mathcal{D}(p) = \mathcal{D}\left(\frac{1}{k(p)}\right)$, where $k(p) = \max\{|k_1|, ..., |k_n|\}$, and where $k_1, ..., k_n$ are the principal curvatures of Σ^n, at the point $p \in \Sigma^n$.*

In particular, if Σ^n is compact, then there exists a sampling of Σ^n having uniformly bounded density. Note, however, that this is not necessarily the optimal scheme (see [12]).

The constructive proof of this theorem is based on the existence of the so-called *fat* (or *thick*) triangulations (see [11]). The density of the vertices of the triangulation (i.e. of the sampling) is given by the inverse of the maximal principal curvature. An essential step in the construction of the said triangulations consists of isometrically embedding of Σ^n in some $\mathbb{R}^N$, for large enough N (see [10]), where the existence of such an embedding is guaranteed by Nash's Theorem ([9]). Resorting to such a powerful tool as Nash's Embedding Theorem appears to be an impediment of our method, since the provided embedding dimension N is excessively high (even after further refinements due to Gromov [4] and Günther [5]). Furthermore, even finding the precise embedding dimension (lower than the canonical N) is very difficult even for simple manifolds. However, as we shall indicate in the next section, this high embedding dimension actually becomes an advantage, at least from the viewpoint of information theory.

The resultant sampling scheme is in accord with the classical Shannon theorem, at least for the large class of (bandlimited) signals that also satisfy the condition of being $\mathcal{C}^2$ curves. In our proposed geometric approach, the radius of curvature substitutes for the condition of the Nyquist rate. To be more precise, our approach parallels, in a geometric setting, the *local bandwidth* of [7] and [16]. In other words, manifolds with bounded curvature represent a generalization of the *locally band limited signals* considered in those papers.

We concentrate here only on some of the consequences of Theorem 1. More precisely, we present, in Sections 2 and 3, two applications of our geometric sampling method and of the embedding technique employed in the proof, namely to the vector quantization of images and to determining the embedding dimension in Zador's Theorem, respectively. Further directions of study are briefly discussed in the concluding section.

2. Vector Quantization for Images

A complementary byproduct of the constructive proof of Theorem 1 is a precise method of *vector quantization* (or *block coding*). Indeed, the proof of Theorem 1 consists in the construction of a Voronoi (Dirichlet) cell complex $\{\bar{\gamma}_k^n\}$ (whose vertices will provide the sampling points). The centers a_k of the cells (satisfying a certain geometric density condition) represent, as usual, the *decision vectors*. An advantage of this approach, besides its simplicity, is entailed by the possibility to estimate the error in terms of length and angle distortion when passing from the cell complex $\{\bar{\gamma}_k^n\}$ to the Euclidean cell complex $\{\bar{c}_k^n\}$ having the same set of vertices as $\{\bar{\gamma}_k^n\}$ (see [10]). Indeed, in contrast to other related studies, our method not only produces a piecewise-flat simplicial approximation of the given manifold, it also actually renders a simplicial complex on the manifold. Moreover, one can actually compute the local distortion resulting by passing from the Euclidean geometry of the piecewise-flat approximation to the intrinsic geometry of its projection on the manifold. If $M = M^n$ is a manifold without boundary, then locally, for any triangulation patch the following inequality holds [10]:

$$\frac{3}{4} d_M(x,y) \leq d_{eucl}(\bar{x},\bar{y}) \leq \frac{5}{3} d_M(x,y)\,;$$

where d_{eucl}, d_M denote the Euclidean and intrinsic metric (on M) respectively, and where $x, y \in M$ and $\bar{x}, \bar{y}$ are their preimages on the piecewise-flat complex. For manifolds with boundary, the same estimate holds (for the $intM$ and ∂M), except for a (small) zone of "mashing" triangulations (see [11]), where the following weaker distortion formula is easily obtained:

$$\frac{3}{4} d_M(x,y) - f(\theta)\eta_\partial \leq d_{eucl}(\bar{x},\bar{y}) \leq \frac{5}{3} d_M(x,y) + f(\theta)\eta_\partial\,;$$

where $f(\theta)$ is a constant depending on the $\theta = \min\{\theta_\partial, \theta_{\mathrm{int}\,M}\}$ – the fatness of the triangulation of ∂M and $\mathrm{int}\,\mathrm{M}$, respectively, and η_∂ denotes the *mesh* of the triangulation of a certain neighbourhood of ∂M (see [11]). In other words, the (local) projection mapping π between the triangulated manifold M and its piecewise-flat approximation Σ is (locally) *bi-lipschitz* if M is open, but only a *quasi-isometry* (or *coarsely bi-lipschitz*) if the boundary of M is not empty.

But the main advantage of a geometric sampling of images resides in the fact that the sampling is done according to the geometric, hence intrinsic, features of the image, rather in the arbitrary (as far as features are concerned) manner of classical approach that transforms the image into a 1-dimensional array (signal). Therefore, the resulting sampling is adaptive, hence sparse in regions of low curvature, and, as shown in [1], it is even compressive in some special cases.

3. Zador's Theorem

A more important application stems, however, from Zador's Theorem [15], implying that we can turn into an advantage the inherent "curse of dimensionality". Indeed, by of Zador's Theorem, the *average mean squared error per dimension*:

$$\mathcal{E} = \frac{1}{N} \int_{\mathbb{R}^N} d_{eucl}(x, p_i) p(x) dx\,,$$

p_i being the *code point* closest to x and $p(x)$ denoting the *probability density function* of x, can be reduced by making avail of higher dimensional quantizers (see [2]). Since for embedded manifolds it obviously holds that $p(x) = p_1(x)\chi_M$, we obtain:

$$\mathcal{E} = \frac{1}{N} \int_{M^n} d_{eucl}(x, p_i) p_1(x) dx\,,$$

It follows that, if the main issue is accuracy, not simplicity, then 1-dimensional coding algorithms (such as the classical Ziv-Lempel algorithm) perform far worse than higher dimensional ones. Of course, there exists an upper limit for the coding dimension, since otherwise one could just code the whole data as one N-dimensional vector (albeit of unpractically high dimension). The geometric coding method proposed here provides a *natural* high dimension for the quantization of M^n – the embedding dimension N. Moreover, it closes (at least for images and any other data that can be represented as Riemannian manifolds) an open problem related to Zador's Theorem: finding a constructive method to determine the dimension of the quantizers (Zador's proof is nonconstructive). In fact, for a uniformly distributed input (as manifolds, hence noiseless images, can assumed to be, at least in first approximation) a better estimate of the average mean squared error per dimension can be obtained, namely:

$$\mathcal{E} = \frac{\frac{1}{N}\int_{M^n} d_{eucl}(x,p_i)dx}{\int_{M^n} dx} = \frac{\frac{1}{N}\int_{M^n} d_{eucl}(x,p_i)dx}{\mathcal{V}_n(M^n)dx}\,,$$

where $\mathcal{V}_n$ denotes the n-dimensional volume (area) of M. Whence, for compact manifolds one obtains the following expression for $\mathcal{E}$:

$$\mathcal{E} = \frac{\frac{1}{N}\int_{M^n} d_{eucl}(x,p_i)dx}{\sum_i^m \int_{V_i} dx} = \frac{\frac{1}{N}\int_{M^n} d_{eucl}(x,p_i)dx}{\sum_i^m \mathcal{V}_n(V_i)dx}\,,$$

where V_i represent the Voronoi cells of the partition.

Moreover, we have the following estimate for the *quantizer problem*, that is: Chose centers of cells such that the quantity

$$\mathcal{Q} = \frac{1}{N} \frac{\frac{1}{m}\int_{M^n} d_{eucl}(x,p_i)dx}{\left(\frac{1}{m}\sum_i^m \mathcal{V}_n\right)^{1+\frac{2}{N}}}\,.$$

is minimized. Here, again, the high embedding dimension N furnishes us with yet an additional advantage. Indeed, manifolds N increases dramatically, even for compact manifolds and even taking into consideration Gromov's and Günther's improvement of Nash's original method (see [4], resp. [5]). For instance, $n = 2$ requires embedding dimension $N = 10$ and $n = 3$ the necessitates $N = 14$. Hence, for large enough n one can write the following rough estimate:

$$\mathcal{Q} \approx \frac{1}{N} \frac{\int_{M^n} d_{eucl}(x, p_i) dx}{\sum_i^m \mathcal{V}_n} .$$

4. Conclusions and Future work

As we have stressed above, our geometrical approach to sampling lends itself to consideration of a much broader range of topics in communications, for such problems as Coding, Channel Capacity, amongst others (see [13]). In particular, and almost as an afterthought of the ideas presented in Section 2, it offers a new method for PCM (*pulse code modulation* – see [2] for a brief yet lucid presentation) of images, considered as such and not as 1-dimensional signals. This approach is endowed with an inherent advantage in that the sampling points are associated with relevant geometric features (via curvature) of the image, viewed as a manifold of dimension ≥ 2, and are not chosen via the Nyquist rate of some rather arbitrarily computed 1-dimensional signal. Moreover, the sampling is in this case adaptive and, indeed, compressive, lending itself to interesting technological benefits.
The implementation of the PCM method described above, as well as experimenting with the geometric quantization method, represent the applicative directions of study that are natural and interesting to pursue further. A better understanding of the geometry of images, included color, texture and other relevant features, in terms of curvature, represent the theoretical directions to be pursued in future. In particular, determining the lowest embedding dimension and finding global curvature constraints are, as we have seen, important for a highly compressive sampling.

5. The role of curvature

We briefly discuss here the crucial role of curvature in determining the embedding dimension (and hence the Zador dimension) by illustrating it on a "toy" example, namely that of the torus.
For a "round" torus of revolution T_r^2 in $\mathbb{R}^3$, the embedding dimension is $N = 3$, since the metric of T_r^2 is the intrinsic one induced by the Euclidian one of the ambient space $\mathbb{R}^3$, thus in this case our method does not depart too much from standard ones. However, if one considers the *flat* torus T_f^2, i.e. of Gaussian curvature $K \equiv 0$, then the minimal dimension needed for isometric embedding is $N = 4$ (see, e.g. [3]). (Before we proceed further, let us note that such tori arise naturally when considering planar rectangles with opposite sides identified – that is, "glued" – via translations. In a practical context, these would model 2-dimensional repetitive patterns on a computer screen, e.g. screen savers. Flat tori also appear in another context relevant to Computer Graphics and Image Processing, namely as solutions for discrete curvature flows (on triangular meshes), see e.g. [8].) In general, given a 2-dimensional torus, equipped with generic Riemannian metric, the whole range of dimensions, up to, and including, the one prescribed by the Nash-Gromov-Günther Theorem, is possible. There are huge differences arising not only from the sign of the curvature, but from its "speed of change" as well – for a exhaustive treatment of this subject see [6].

6. Acknowledgments

The authors would like to thank Professor Peter Maass, for his constructive critique and encouragement. The first author would also like to thank Professor Shahar Mendelson – his warm support is gratefully acknowledged.

References:

[1] Eli Appleboim, Emil Saucan and Yehoshua Y. Zeevi. Geometric Sampling For Signals With Applications to Images. *Proceedings of Sampta 2007*, 2008.

[2] John H. Conway and Neil J. A. Sloane *Sphere Packings, Lattices and Groups*. Springer, New York, 1999.

[3] Manfredo P. do Carmo *Differential Geometry of Curves and Surfaces*. Prentice-Hall, Englewood Cliffs, N.J., 1976.

[4] Mikhail Gromov. *Partial differential relations*, Springer-Verlag, Ergeb. der Math. 3 Folge, Bd. 9, Berlin-Heidelberg-New-York, 1986.

[5] Matthias Günther. Isometric embeddings of Riemannian manifolds. *Proc. ICM Kyoto*, pages 1137–1143, 1990.

[6] Qing Han and Jia-Xing Hong Isometric embeddings of Riemannian manifolds in Euclidean Spaces. AMS MSM 130, Providnce, RI, 2006.

[7] K. Horiuchi. Sampling principle for continuous signals with time-varying bands. *Information and Control*, 13(1): 53-61, 1968.

[8] Miao Jin, J. Kim and David Gu. Discrete Surface Ricci Flow: Theory and Applications. In *Mathematics of Surfaces*, LNCS 4647, pages 209–232, 2007.

[9] John Nash. The embedding problem for Riemannian manifolds. *Ann. of Math.* 63:20–63, 1956.

[10] Kirsi Peltonen. On the existence of quasiregular mappings. *Ann. Acad. Sci. Fenn., Series I Math., Dissertationes* 1992.

[11] Emil Saucan. Note on a theorem of Munkres. *Mediterr. j. math.* 2(2):215–229, 2005.

[12] Emil Saucan, Eli Appleboim, and Yehoshua Y Zeevi. Sampling and Reconstruction of Surfaces and Higher Dimensional Manifolds. *Journal of Mathematical Imaging and Vision* 30(1):105–123, 2008.

[13] Emil Saucan, Eli Appleboim, and Yehoshua Y Zeevi. Geometric Approach to Sampling and Communication. Technion CCIT Report #707, November 2008.

[14] Claude E. Shannon. Communication in the presence of noise. *Proceedings of the IRE* 37(1):10–21, 1949.

[15] Paul G. Zador. Asymptotic Quantization Error of Continuous Signals and the Quantization Dimension. *IEEE Trans. on Info. Theory*, 12(1):23–86, 1982.

[16] Yehoshua Y. Zeevi and E. Shlomot. Nonuniform sampling and antialiasing in image representation. *IEEE Trans. Signal Process.*, 41(3):1223–1236, 1993.

On average sampling restoration of Piranashvili–type harmonizable processes

Andriy Ya. Olenko† and Tibor K. Pogány‡

† Department of Mathematics and Statistics, La Trobe University, Victoria 3086, Australia.
‡ Faculty of Maritime Studies, University of Rijeka, Studentska 2, HR-51000 Rijeka, Croatia.
a.olenko@latrobe.edu.au, poganj@pfri.hr

Abstract:

The harmonizable Piranashvili – type stochastic processes are approximated by a finite time shifted average sampling sum. Truncation error upper bound is established; various consequences and special cases are discussed.

MSC(2000): 42C15, 60G12, 94A20.

Keywords: WKS sampling theorem; time shifted sampling; Piranashvili–, Loève–, Karhunen– harmonizable stochastic process; weakly stationary stochastic process; local averages; average sampling reconstruction.

1. Introduction and preparation

Given a probability space $\big(\Omega, \mathfrak{F}, \mathsf{P}\big)$ and the related Hilbert–space $L_2(\Omega) := \{X\colon \mathsf{E}|X|^2 < \infty\}$. Let us consider a non–stationary, centered stochastic $L_2(\Omega)$–process $\xi\colon \mathbb{R}\times\Omega \mapsto \mathbb{R}$ having covariance function (associated to some domain $\Lambda \subseteq \mathbb{R}$ with some sigma–algebra $\sigma(\Lambda)$) in the form:

$$B(t,s) = \int_\Lambda\int_\Lambda f(t,\lambda) f^*(s,\mu) F_\xi(\mathrm{d}\lambda, \mathrm{d}\mu), \qquad (1)$$

with analytical exponentially bounded kernel function $f(t,\lambda)$, while F_ξ is a positive definite measure on $\mathbb{R}^2$ provided the total variation $\|F_\xi\|(\Lambda,\Lambda)$ of the spectral distribution function F_ξ such that satisfies

$$\|F_\xi\|(\Lambda,\Lambda) = \int_\Lambda\int_\Lambda \big|F_\xi(\mathrm{d}\lambda, \mathrm{d}\mu)\big| < \infty.$$

(We mention that the sample function $\xi(t) \equiv \xi(t,\omega_0)$ and $f(t,\lambda)$ possess the same exponential types [1, Theorem 4], [11, Theorem 3]). Then, by the Karhunen–Cramér theorem the process $\xi(t)$ has the spectral representation as a Lebesgue integral

$$\xi(t) = \int_\Lambda f(t,\lambda) Z_\xi(\mathrm{d}\lambda); \qquad (2)$$

in (1) and (2)

$$F_\xi(S_1,S_2) = \mathsf{E} Z_\xi(S_1) Z_\xi^*(S_2) \qquad S_1, S_2 \subseteq \sigma(\Lambda).$$

Such a process will be called Piranashvili process in the sequel [11], [12].

Being $f(t,\lambda)$ entire, it possesses the Maclaurin expansion $f(t,\lambda) = \sum_{n=0}^\infty f^{(n)}(0,\lambda)t^n/n!$. Put

$$\gamma := \sup_\Lambda c(\lambda) = \sup_\Lambda \overline{\lim_n} \sqrt[n]{|f^{(n)}(0,\lambda)|} < \infty\,. \qquad (3)$$

As the exponential type of $f(t,\lambda)$ is equal to γ, for all $w > \gamma$ there holds

$$\xi(t) = \sum_{n\in\mathbb{Z}} \xi\Big(\frac{n\pi}{w}\Big) \frac{\sin(wt - n\pi)}{wt - n\pi}, \qquad (4)$$

uniformly in the mean square and in the almost sure sense [11, Theorem 1]. This result we call Whittaker–Kotel'nikov–Shannon (WKS) stochastic sampling theorem [12].

Specifying $F_\xi(x,y) = \delta_{xy} F_\xi(x)$ in (1) we conclude the Karhunen–representation of the covariance function

$$B(t,s) = \int_\Lambda f(t,\lambda) f^*(s,\lambda) F_\xi(\mathrm{d}\lambda).$$

Also, putting $f(t,\lambda) = e^{it\lambda}$ in (1) one gets the Loève-representation:

$$B(t,s) = \int_\Lambda\int_\Lambda e^{i(t\lambda - s\mu)} F_\xi(\mathrm{d}\lambda, \mathrm{d}\mu).$$

Here is $c(\lambda) = |\lambda|$. Therefore, WKS–formula (4) holds for all $w > \gamma = \sup|\Lambda|$. Then, the Karhunen process with the Fourier kernel $f(t,\lambda) = e^{it\lambda}$ we recognize as the weakly stationary stochastic process having covariance

$$B(\tau) = \int_\Lambda e^{i\tau\lambda} F_\xi(\mathrm{d}\lambda), \qquad \tau = t - s.$$

Deeper insight into different kind harmonizabilities present [5, 13, 14] and the related references therein. Finally, using $\Lambda = [-w, w]$ for some finite w in this considerations, we get the band–limited variants of the same kind processes.

By physical and applications reasons the measured samples in practice may not be the exact values of the measured process $\xi(t)$, or its covariance $B(t,s)$ itself, near to the sample time t_n, but only the local average of the signal ξ near to t_n. So, the measured sample values will be

$$\langle \xi, u_n\rangle_U = \int_U \xi(x) u_n(x)\mathrm{d}x, \quad U = \mathrm{supp}(u_n) \qquad (5)$$

for a sequence $\mathbf{u} := \big(u_n(t)\big)_{n\in\mathbb{Z}}$ of non–negative, normalized, that is $\langle 1, u_n\rangle \equiv 1$, averaging functions such that

$$\mathrm{supp}(u_n) \subseteq \big[t_n - \sigma_n', t_n + \sigma_n''\big]\,. \tag{6}$$

The local averaging method was introduced by Gröchenig [2] and developed by Butzer and Lei. Recently Sun and Zhou gave some results in this direction, while the stochastic counterpart of this average sampling was intensively studied in the last three–four years by He, Song, Sun, Yang and Zhu in a set of articles [15], [16] and their references therein; see for example the exhaustive references list in [4]. The listed, recently considered stochastic average sampling results are restricted to weakly stationary stochastic processes, while the approximation average sampling sums are used around the origin.

Our intentions are to extend these results to time shifted average sampling, considered for the very wide class of Piranashvili processes.

2. Time shifted average sampling

Now, instead to follow the approach used in [16] we take time shifted [7], [8] finite average sampling sum in approximating the initial stochastic signal ξ. First, we consider weighted average over $\mathfrak{J}_n(t) := \big[n\pi/w - \sigma_n'(t), n\pi/w + \sigma_n''(t)\big]$ for the measured value of $\xi(t)$ at $n\pi/w$, $n \in \mathbb{I}_N(t)$ where

$$\mathbb{I}_N(t) := \{n \in \mathbb{Z}\colon |tw/\pi - n| \le N\}, \qquad N \in \mathbb{N}.$$

Let N_t be the integer nearest to tw/π.

By obvious reasons we restrict the study to

$$\sigma := \max_{\mathbb{I}_N(t)} \sup_{\mathbb{R}} \max\big(\sigma_n'(t), \sigma_n''(t)\big) \le \frac{\pi}{2w}\,.$$

Let us define the time shifted average sampling approximation sum in the form

$$\mathcal{A}_{\mathbf{u}}(\xi;t) = \sum_{\mathbb{Z}} \langle \xi, u_n\rangle_{\mathfrak{J}_n(t)} \cdot \frac{\sin(wt - n\pi)}{wt - n\pi}\,,$$

and its truncated variant

$$\mathcal{A}_{\mathbf{u},N}(\xi;t) = \sum_{\mathbb{I}_N(t)} \langle \xi, u_n\rangle_{\mathfrak{J}_n(t)} \cdot \frac{\sin(wt - n\pi)}{wt - n\pi}\,.$$

One defines mean–square, time shifted, average sampling truncation error $\mathfrak{T}_{\mathbf{u},N}(\xi;t) := \mathsf{E}\big|\xi(t) - \mathcal{A}_{\mathbf{u},N}(\xi;t)\big|^2$. Now, we are interested in some reasonably simple efficient mean square truncation error upper bound appearing in the approximation $\xi(t) \approx \mathcal{A}_{\mathbf{u},N}(\xi;t)$.

Let us introduce some auxiliary results. As N_x stands for the integer nearest to xw/π, $x \in \mathbb{R}$, let

$$\Gamma_N(x) := \Big\{z \in \mathbb{C}\colon |z - N_x| \le \big(N + \tfrac{1}{2}\big)\tfrac{\pi}{w}\Big\}, \quad N \in \mathbb{N}.$$

In what follows denote $\mathrm{int}(R)$ the interior of some R, while the series

$$\lambda(q) := \sum_{n=1}^{\infty} \frac{1}{(2n-1)^q}$$

stands for the Dirichlet lambda function.

Theorem 1 *Let $f(z)$ be entire, bounded on the real axis and exponentially bounded having type $\gamma < w$. Denote*

$$L_f := \sup_{\mathbb{R}} \big|f(x)\big|, \quad L_0(z) := \frac{2wL_f|\sin(wz)|}{\pi(w-\gamma)\big(1 - e^{-\pi}\big)}\,.$$

Then for all $z \in \mathrm{int}\big(\Gamma_N(x)\big)$ and $N \in \mathbb{N}$ enough large it holds

$$\Big|\sum_{\mathbb{Z}\setminus\mathbb{I}_N(x)} f\Big(n\frac{\pi}{w}\Big)\frac{\sin(wz - n\pi)}{wz - n\pi}\Big| \le \frac{L_0(z)e^{-(N+1/2)\pi(w-\gamma)/w}}{(N+1/2)\big|1 - \frac{|z-N_x|w}{(N+1/2)\pi}\big|} < \frac{L_0(z)}{N}\,. \tag{7}$$

The proving method is contour integration, following Piranashvili's traces [11]. Denote here and in what follows

$$Y_N(\xi;t) := \sum_{\mathbb{I}_N(t)} \xi\Big(\frac{n\pi}{w}\Big)\frac{\sin(wt - n\pi)}{wt - n\pi}$$

the time shifted truncated WKS restoration sum.

By simple use of (1), (2) and the Theorem 1 one deduces the following modest generalization of [11, Theorem 2] to time shifted case of sampling restoration procedure.

Theorem 2 *Let $\xi(t)$ be a Piranashvili process with exponentially bounded kernel function $f(t,\lambda)$ and let*

$$\widetilde{L}_f := \sup_{\mathbb{R}}\sup_{\Lambda}|f(t,\lambda)|,$$
$$\widetilde{L}_0(t) := \frac{2\widetilde{L}_f w\,|\sin(wt)|}{\pi(w-\gamma)\big(1-e^{-\pi}\big)}\,. \tag{8}$$

Then for all $t \in \mathrm{int}\big(\Gamma_N(t)\big)$, we have

$$\mathsf{E}\big|\xi(t) - Y_N(\xi;t)\big|^2 < \frac{\widetilde{L}_0^2(t)}{N^2}\,\|F_\xi\|(\Lambda,\Lambda)\,. \tag{9}$$

Remark 1 *Let us point out that the straightforward consequence of* (9) *is not only the exact L_2–restoration of the initial Piranashvili–type harmonizable process ξ by a sequence of approximants $Y_N(\xi;t)$ when $N \to \infty$, but since*

$$\mathsf{E}\big|\xi(t) - Y_N(\xi;t)\big|^2 = \mathcal{O}\big(N^{-2}\big)\,,$$

the perfect reconstruction is possible in the a.s. *sense as well (by the celebrated Borel–Cantelli Lemma).*

Second, the first order difference $\Delta_{x,y}B$ [3] of $B(t,s)$ on the plane satisfies

$$\begin{aligned}\big(\Delta_{x,y}B\big)(t,s) &= B(t+x,s+y) - B(t+x,s)\\ &\quad - B(t,s+y) + B(t,s)\\ &= \int_0^x\int_0^y \frac{\partial^2}{\partial u\partial v}B\big(t+u,s+v\big)\,\mathrm{d}v\mathrm{d}u\,.\end{aligned} \tag{10}$$

Theorem 3 *Let $\xi(t)$ be a Piranashvili process with the covariance $B(t,t) \in C^2(\mathbb{R})$. Let (p,q) be a conjugated Hölder pair of exponents:*

$$\frac{1}{p} + \frac{1}{q} = 1, \qquad p > 1\,.$$

Then we have

$$\mathsf{E}\big|Y_N(\xi;t)-\mathcal{A}_{\mathbf{u},N}(\xi;t)\big|^2 \le \frac{\mathsf{C}_q\,\pi^2}{4w^2}\sup_{\mathbb{R}}\big|B''(t,t)\big|\cdot(2N+1)^{2/p}\,, \qquad (11)$$

where

$$\mathsf{C}_q=\Big(1+\frac{2^{q+1}\,|\sin(wt)|^q}{\pi^q}\,\lambda(q)\Big)^{2/q}. \qquad (12)$$

PROOF. Having on mind (1), the properties of averaging functions sequence $\mathbf{u}$ and (10), we clearly derive

$$\begin{aligned}
&\mathsf{E}\big|Y_N(\xi;t)-\mathcal{A}_{\mathbf{u},N}(\xi;t)\big|^2\\
&=\mathsf{E}\Big|\sum_{\mathbb{I}_N(t)}\langle\xi\big(\tfrac{n\pi}{w}\big)-\xi(x),u_n\rangle_{\mathfrak{I}_n(t)}\cdot\frac{\sin(wt-n\pi)}{wt-n\pi}\Big|^2\\
&=\sum_{\mathbb{I}_N^2(t)}\int_{-\sigma_n'(t)}^{\sigma_n''(t)}\int_{-\sigma_m'(t)}^{\sigma_m''(t)}u_n(x+n\tfrac{\pi}{w})u_m(y+m\tfrac{\pi}{w})\\
&\quad\cdot\int_0^x\int_0^y\frac{\partial^2}{\partial u\partial v}B\big(u+n\tfrac{\pi}{w},v+m\tfrac{\pi}{v}\big)\mathrm{d}v\mathrm{d}u\\
&\quad\cdot\frac{\sin(wt-n\pi)}{wt-n\pi}\,\frac{\sin(wt-m\pi)}{wt-m\pi}\\
&\le\sum_{\mathbb{I}_N^2(t)}\Big|\frac{\sin(wt-n\pi)}{wt-n\pi}\Big|\,\Big|\frac{\sin(wt-m\pi)}{wt-m\pi}\Big|\\
&\quad\cdot\sup_{x,y\le\sigma}\Big|\int_0^x\int_0^y\frac{\partial^2}{\partial u\partial v}B\big(u+n\tfrac{\pi}{w},v+m\tfrac{\pi}{v}\big)\mathrm{d}v\mathrm{d}u\Big|
\end{aligned}$$

being $\mathbf{u}$ normalized. For the sake of brevity let us denote $H_\sigma(n,m)$ the sup–term in the last display. Then, by the Hölder inequality with conjugate exponents $p,q;\ p>1$, we get

$$\mathsf{E}\big|Y_N(\xi;t)-\mathcal{A}_{\mathbf{u},N}(\xi;t)\big|^2 \le\Big\{\sum_{\mathbb{I}_N^2(t)}H_\sigma^p(n,m)\Big\}^{1/p}\Big\{\sum_{\mathbb{I}_N(t)}\Big|\frac{\sin(wt-n\pi)}{wt-n\pi}\Big|^q\Big\}^{2/q}.$$

It is not hard to see that for all $n,m\in\mathbb{I}_N(t)$ there holds

$$\begin{aligned}
H_\sigma(n,m)&\le\sigma^2\sup_{\mathbb{R}^2}\Big|\frac{\partial^2B(t,s)}{\partial t\partial s}\Big|\\
&\le\frac{\pi^2}{4w^2}\sup_{\mathbb{R}^2}\Big|\frac{\partial^2B(t,s)}{\partial t\partial s}\Big|\,.
\end{aligned}$$

Applying now the Cauchy–Bunyakovsky–Schwarz inequality to the covariance ∂^2B, we deduce

$$\begin{aligned}
\sup_{\mathbb{R}^2}\Big|\frac{\partial^2B(t,s)}{\partial t\partial s}\Big|&\le\sup_{\mathbb{R}}\Big|\frac{\partial^2B(t,t)}{\partial t^2}\Big|\\
&=\sup_{\mathbb{R}}|B''(t,t)|\,.
\end{aligned}$$

It remains to evaluate the sum of qth power of the sinc–functions. As

$$\frac{\sin(wt-N_t\pi)}{wt-N_t\pi}\le1$$

we conclude

$$\begin{aligned}
\sum_{\mathbb{I}_N(t)}\Big|\frac{\sin(wt-n\pi)}{wt-n\pi}\Big|^q&\le1+C\sum_{n=1}^N\Big\{\frac{1}{(n-\Delta)^q}+\frac{1}{(n+\Delta)^q}\Big\}\\
&<1+2C\sum_{n=1}^\infty\frac{1}{(n-1/2)^q}\\
&<1+2^{q+1}C\,\lambda(q)\,,
\end{aligned}$$

where

$$C=\frac{|\sin(wt)|^q}{\pi^q}\,.$$

Collecting all these estimates, we deduce (11). □

3. Main result

We are ready to formulate our upper bound result for the mean square, time shifted average sampling truncation error $\mathfrak{T}_{\mathbf{u},N}(\xi;t)$. The almost sure sense restoration procedure has been treated too.

As we use average sampling sum $\mathcal{A}_{\mathbf{u},N}(\xi;t)$ instead of $Y_N(\xi;t)$ to obtain asymptotically vanishing $\mathfrak{T}_{\mathbf{u},N}(\xi;t)$, it is not enough letting $N\to\infty$ as in Remark 1. For average sampling we need additional conditions upon w or σ to guarantee smaller average intervals for larger/denser sampling grids.

Theorem 4 *Assume the conditions of* Theorems 2 *and* 3 *have been fulfilled. Then, we have*

$$\mathfrak{T}_{\mathbf{u},N}(\xi;t)\le\frac{2\widetilde{L}_0^2(t)}{N^2}\,\|F_\xi\|(\Lambda,\Lambda) +\frac{\mathsf{C}_q\,\pi^2}{2w^2}\sup_{\mathbb{R}}\big|B''(t,t)\big|\cdot(2N+1)^{2/p}\,, \qquad (13)$$

where $\widetilde{L}_0$, C_q are described by (8), (11) *respectively. Moreover, when $w=\mathcal{O}\big(N^{1/2+1/p+\varepsilon}\big)$, $\varepsilon>0$, we have*

$$\mathsf{P}\big\{\lim_{N\to\infty}\mathcal{A}_{\mathbf{u},N}(\xi;t)=\xi(t)\big\}=1 \qquad (14)$$

for all $t\in\mathbb{R}$.

PROOF. By direct calculation we deduce

$$\begin{aligned}
\mathfrak{T}_{\mathbf{u},N}(\xi;t)&=\mathsf{E}\big|\xi(t)-\mathcal{A}_{\mathbf{u},N}(\xi;t)\big|^2\\
&=\mathsf{E}\big|\xi(t)-Y_N(\xi;t)+Y_N(\xi;t)-\mathcal{A}_{\mathbf{u},N}(\xi;t)\big|^2\\
&\le2\mathsf{E}\big|\xi(t)-Y_N(\xi;t)\big|^2\\
&\quad+2\mathsf{E}\big|Y_N(\xi;t)-\mathcal{A}_{\mathbf{u},N}(\xi;t)\big|^2\,.
\end{aligned}$$

Now, we get the asserted upper bound by (9) and (11).

To derive (14), we apply the Chebyshev inequality to evaluate the probability

$$\mathsf{P}_N:=\mathsf{P}\big\{\big|\xi(t)-\mathcal{A}_{\mathbf{u},N}(\xi;t)\big|\ge\eta\big\}\le\eta^{-2}\mathfrak{T}_{\mathbf{u},N}(\xi;t)\,.$$

Accordingly, since $\widetilde{L}_0(t)=\mathcal{O}(1)$ as $N\to\infty$, we have

$$\sum_{\mathbb{N}}\mathsf{P}_N\le K\sum_{\mathbb{N}}\Big(\frac{1}{N^2}+\frac{(2N+1)^{2/p}}{w^2}\Big)<\infty,$$

K being a suitable absolute constant. Therefore, by the Borel–Cantelli Lemma, the the a.s. convergence result (14) holds true. □

Remark 2 Theorem 4 *ensures the perfect time shifted average sampling restoration in the mean square sense when* $w = \mathcal{O}\big(N^{1/p+\varepsilon}\big)$, $\varepsilon > 0$*:*

$$\lim_{N\to\infty} \mathfrak{T}_{\mathbf{u},N}(\xi;t) = 0\,.$$

The a.s. sense restoration (14) *requires stronger assumption, it holds when* $w = \mathcal{O}\big(N^{1/2+1/p+\varepsilon}\big)$.

Remark 3 *In both cases we use the so called* approximate sampling procedure*, that is, when in the restoration procedure* $w \to \infty$ *in some fashion. The consequence of these results is that we have to restrict ourselves to the case* $\Lambda = \mathbb{R}$*, such that we recognize as the non–bandlimited Piranashvili type harmonizable process case.*

The importance of approximate sampling procedures for investigations of aliasing errors in sampling restorations and different conditions on joint asymptotic behaviour of N *and* w *have been discussed in detail in* [7].

4. Conclusions

We have analyzed upper bounds on truncation error for time shifted average sampling restorations in the stochastic initial signal case. The convergence of the truncation error to zero was discussed. However, certain new questions immediately arise:

- to derive sharp upper bounds in Theorems 3 and 4;
- to obtain new results for L_p–processes using recent deterministic findings [9], [10];
- to obtain similar results for irregular/nonuniform sampling restoration using methods exposed in [6] and [10].

Acknowledgements

The recent investigation was supported in part by Research Project No. 112-2352818-2814 of Ministry of Sciences, Education and Sports of Croatia and in part by La Trobe University Research Grant–501821 "Sampling, wavelets and optimal stochastic modelling".

References:

[1] Yuri K. Belyaev. Analytical random processes. *Teor. Verojat. Primenen.* **IV**(4): 437–444, 1959. (in Russian)

[2] Karlheinz Gröchenig. Reconstruction algorithms in irregular sampling. *Math. Comput.* **59**: 181–194, 1992.

[3] Muhammed K. Habib and Stamatis Cambanis. Sampling approximation for non–band–limited harmonizable random signals. *Inform. Sci* **23**:143–152, 1981.

[4] Gaiyun He, Zhanjie Song, Deyun Yang and Jianhua Zhu. Truncation error estimate on random signals by local average. In Y. Shi *et al.*, editors. *ICCS* 2007, Part II, Lecture Notes in Computer Sciences 4488, pages 1075–1082, 2007.

[5] Yûichirô Kakihara. Multidimensional Second Order Stochastic Processes. World Scientific, Singapore, 1997.

[6] Andriy Ya. Olenko and Tibor K. Pogány. Direct Lagrange–Yen type interpolation of random fields. *Theor. Stoch. Proc.* **9(25)**(3–4): 242–254, 2003.

[7] Andriy Ya. Olenko and Tibor K. Pogány. Time shifted aliasing error upper bounds for truncated sampling cardinal series. *J. Math. Anal. Appl.* **324**(1): 262–280, 2006.

[8] Andriy Ya. Olenko and Tibor K. Pogány. On sharp bounds for remainders in multidimensional sampling theorem. *Sampl. Theory Signal Image Process.* **6**(3): 249–272, 2007.

[9] Andriy Ya. Olenko and Tibor K. Pogány. Universal truncation error upper bounds in sampling restoration. (to appear)

[10] Andriy Ya. Olenko and Tibor K. Pogány. Universal truncation error upper bounds in irregular sampling restoration. (to appear)

[11] Zurab A. Piranashvili. On the problem of interpolation of random processes. *Teor. Verojat. Primenen.* **XII**(4): 708–717, 1967. (in Russian)

[12] Tibor K. Pogány. Almost sure sampling restoration of bandlimited stochastic signals. In John R. Higgins and Rudolf L. Stens, editors. *Sampling Theory in Fourier and Signal Analysis: Advanced Topics*, Oxford University Press, pages 203–232, 284–286, 1999.

[13] Maurice B. Priestley. Non–linear and Non–stationary Time Series. Academic Press, London, New York, 1988.

[14] Malempati M. Rao. Harmonizable processes: structure theory. *Einseign. Math. (2)* **28**(3–4): 295–351, 1982.

[15] Zhanjie Song, Zingwei Zhu and Gaizun He. Error estimate on non–bandlimited random signals by local averages. In V.N. Aleksandrov it et al., editors. *ICCS* 2006, Part I, Lecture Notes in Computer Sciences 3991, pages 822–825, 2006.

[16] Zhan–jie Song, Wen–chang Sun, Shou–yuan Yang and Guang–wen Zhu. Approximation of weak sense stationary stochastic processes from local averages. *Sci. China Ser. A* **50**(4): 457–463, 2007.

Sampling of Homogeneous Polynomials

Somantika Datta [(1)], Stephen D. Howard [(2)], and Douglas Cochran [(1)]

(1) Arizona State University, Tempe, Arizona 85287, USA.
(2) Defence Science & Technology Organisation, Edinburgh, South Australia.
somantika.datta@asu.edu, stephen.howard@dsto.defence.au, cochran@asu.edu

Abstract:

Conditions for reconstruction of multivariate homogeneous polynomials from sets of sample values are introduced, together with a frame-based method for explicitly obtaining the polynomial coefficients from the sample data.

1. Introduction

Several authors have noted the importance of interpolation and reconstruction of multivariate polynomials from sample data in applications. Zakhor [10], for example, considered the problem of interpolation of bivariate polynomials from irregularly spaced sample values in connection with two-dimensional filter design and image processing. The case of multivariate polynomials presents significant difficulties not encountered with polynomials of one variable, in particular due to the zeros of these entire functions of several variables not being isolated as occurs in the univariate setting. Consequently, it is not surprising that, in her work, Zakhor develops conditions in which suitable sampling sets are constrained to lie on certain algebraic curves.

Very recent work by Varjú [9] and Benko and Króo [1] develops Weierstraß types of results for approximation of smooth multivariate functions by homogeneous polynomials. This suggests the potential utility of interpolation and reconstruction of homogeneous polynomials from sample values. It is well known that the linear space $H_k(\mathbb{C}^n)$ of homogeneous polynomials of degree k in n complex variables is isomorphic to the space $\mathrm{Sym}^k(\mathbb{C}^n)$ of symmetric k-tensors over $\mathbb{C}^n$. This fact was used by the authors in [3] to develop results concerning frames and grammians on $\mathrm{Sym}^k(\mathbb{C}^n)$. In this paper, a similar perspective is used to derive conditions under which coefficients of a multivariate homogeneous polynomial of known degree can be reconstructed explicitly from sets of sample values. It is shown that a sampling set that suffices for n-variate homogenous polynomials of degree k is also suitable for reconstructing the coefficients of any homogeneous polynomial in n variables of degree $1 \leqslant \ell < k$. Further, it is noted that, modulo general position issues, the number of samples is the crucial issue in determining suitability of a sampling set. Nevertheless, some sampling sets are "better" than others in that they provide snugger frames and hence the numerical advantages they entail. The relative merits of sampling sets in this respect do not depend on the particular polynomial to be reconstructed, thus allowing generically good sampling sets to be designed before any sampling is actually carried out.

Before beginning the mathematical sections of the paper, a few comments on notation and terminology are in order. For $x = [x^{(1)} \ \cdots \ x^{(n)}]^{\mathsf{T}}$ and $y = [y^{(1)} \ \cdots \ y^{(n)}]^{\mathsf{T}}$ in $\mathbb{C}^n$, their inner product will be denoted by

$$\langle x, y \rangle = \sum_{j=1}^{n} \bar{x}^{(j)} y^{(j)}$$

where the bar denotes complex conjugate; i.e., the inner product is conjugate linear in its first argument and linear in its second argument. The corresponding convention will be used for inner products in other complex Hilbert spaces. Given a finite frame $X = \{x_1, ..., x_m\}$ for an n-dimensional complex vector space V, the function $F : V \to \ell_2(\{1, \ldots, m\}) = \mathbb{C}^m$ given by $F(w) = [\langle x_1, w\rangle \ \ldots \ \langle x_m, w\rangle]^{\mathsf{T}}$ will be called the *frame operator* associated with X, while $\mathcal{F} = F^*F : V \to V$ (i.e., the composition of the adjoint of F with F) will be called the *metric operator* associated with X.

The k-fold tensor product $V^{\otimes k}$ of an n-dimensional vector space V is a vector space spanned by elements of the form $v_1 \otimes \cdots \otimes v_k$ where each $v_i \in V$ [8]. The vector $v_1 \otimes \cdots \otimes v_k$ has n^k coordinates $\{v_i^{(\ell)} | i = 1, \ldots, k;\ \ell = 1, \ldots, n\}$ where $v_i^{(\ell)}$ denotes the ℓ^{th} coordinate of the vector v_i. The space of symmetric k-tensors associated with V, denoted $\mathrm{Sym}^k(V)$, is the subspace of $V^{\otimes k}$ consisting of those tensors which remain fixed under permutation (see Chapter 10 of [8]). $\mathrm{Sym}^k(V)$ is spanned by the tensor powers $v^{\otimes k}$ where $v \in V$. If V has dimension n then $\dim \mathrm{Sym}^k(V) = \binom{n+k-1}{k}$. $\mathrm{Sym}^k(V)$ has a natural inner product with the property

$$\left\langle v^{\otimes k}, w^{\otimes k} \right\rangle_{\mathrm{Sym}^k(V)} = \langle v, w \rangle_V^k \,. \tag{1}$$

2. Sampling of Homogeneous Polynomials

It is well known (see, e.g., [8]) that $H_k(\mathbb{C}^n)$, the linear space of homogeneous polynomials of total degree k in variables $\bar{z}^{(1)}, \ldots, \bar{z}^{(n)}$ is isomorphic to $\mathrm{Sym}^k(V)$. This section points out a connection between the condition that $X^{(k)} = \{x_1^{\otimes k}, \ldots, x_m^{\otimes k}\}$ is a frame for $\mathrm{Sym}^k(V)$ and the reconstructability of polynomials in $H_k(\mathbb{C}^n)$ from the values they take at sets of m points in $\mathbb{C}^n$.

Beginning with $k = 1$, let $w \in V = \mathrm{Sym}^1(V)$ and denote by $[w^{(1)} \cdots w^{(n)}]^\mathsf{T} \in \mathbb{C}^n$ the coordinates of w in some orthonormal basis for V. There is an obvious isomorphism that takes $w \in V$ to the polynomial $p_w \in H_1(\mathbb{C}^n)$ defined by $p_w(z^{(1)}, \ldots, z^{(n)}) = w^{(1)}\bar{z}^{(1)} + \cdots w^{(n)}\bar{z}^{(n)}$. If $X = \{x_1, \ldots, x_m\}$ is a frame for V, the associated frame operator $F : V \to \mathbb{C}^m$ is given by

$$F(w) = \begin{bmatrix} \langle x_1, w\rangle \\ \vdots \\ \langle x_m, w\rangle \end{bmatrix} = \begin{bmatrix} p_w(x_1^{(1)}, \ldots, x_1^{(n)}) \\ \vdots \\ p_w(x_m^{(1)}, \ldots, x_m^{(n)}) \end{bmatrix}. \quad (2)$$

In other words, $F(w)$ is a vector of values obtained by evaluating (i.e., "sampling") p_w at the points $x_1, \ldots, x_m$. One may ask whether this set of m sample values is sufficient to uniquely determine p_w.

To address this question, define a sampling function $P_X : H_1 \to \mathbb{C}^m$ by

$$P_X(p) = \begin{bmatrix} p(x_1^{(1)}, \ldots, x_1^{(n)}) \\ \vdots \\ p(x_m^{(1)}, \ldots, x_m^{(n)}) \end{bmatrix}$$

and note that (2) shows the frame operator is given by $F(w) = P_X(p_w)$. Because the frame operator is invertible, w is uniquely determined by $F(w)$. Hence any $p_w \in H_1$ is uniquely determined by its samples $P_X(p_w)$. Conversely, if X fails to frame V, the mapping F defined by (2) is still well-defined, but has non-trivial kernel K. In this case, $P_X(p_w) = P_X(p_{w+u})$ for all $u \in K$. So, in particular, p_w is not uniquely determined from its samples at $x_1, ..., x_m$.

A similar situation occurs for $k > 1$, where the space of interest is $\mathrm{Sym}^k(V)$ and the frame is $X^{(k)} = \{x_1^{\otimes k}, \ldots, x_m^{\otimes k}\}$. As in the $k = 1$ case, mapping a polynomial to its coefficient sequence defines an isomorphism between $H_k(\mathbb{C}^n)$ and $\mathrm{Sym}^k(V)$ for $k > 1$. If $v = w^{\otimes k} \in \mathrm{Sym}^k(V)$ is a pure tensor power of $w \in V$, then

$$\begin{aligned} F^{(k)}(v) &= \begin{bmatrix} \langle x_1^{\otimes k}, w^{\otimes k}\rangle \\ \vdots \\ \langle x_m^{\otimes k}, w^{\otimes k}\rangle \end{bmatrix} \\ &= \begin{bmatrix} \langle x_1, w\rangle^k \\ \vdots \\ \langle x_m, w\rangle^k \end{bmatrix} = \begin{bmatrix} p_v(x_1) \\ \vdots \\ p_v(x_m) \end{bmatrix} \end{aligned}$$

where $p_v \in H_k$ is defined by $p_v(z) = \langle z, w\rangle^k$. $\mathrm{Sym}^k(V)$ is spanned by pure tensor powers of elements in V [8]. Thus, for arbitrary $v \in \mathrm{Sym}^k(V)$, $F^{(k)}(v)$ is a vector of m samples of a polynomial in H_k taken at points $x_1, ..., x_m$. Thus, as in the $k = 1$ case, polynomials in H_k are uniquely determined by the samples $P_X^{(k)}(p) = [p(x_1), \ldots, p(x_m)]^\mathsf{T}$ if and only if $X^{(k)}$ frames $\mathrm{Sym}^k(V)$. Theorem 1 given below implies that if one can reconstruct a polynomial in $H_k(\mathbb{C}^n)$ from a certain sampling set then the same set can be used to reconstruct polynomials in $H_\ell(\mathbb{C}^n)$ for all $1 \leqslant \ell < k$. Conversely, almost every sampling set in $\mathbb{C}^n$ for H_1 gives rise to a sampling set for H_k where $k > 1$, provided there are enough vectors in the set.

Theorem 1. *(i) Given n and m with $m \geqslant n$, if $X^{(k)} = \{x_1^{\otimes k}, x_2^{\otimes k}, \ldots, x_m^{\otimes k}\}$ is a frame for $\mathrm{Sym}^k(V)$, then $X^{(\ell)}$ is a frame for $\mathrm{Sym}^\ell(V)$ for all $1 \leqslant \ell < k$.*
(ii) Almost every set of m vectors in $\mathbb{C}^n$ such that $m \geqslant \binom{n+k-1}{k}$ results in a frame for $\mathrm{Sym}^k(\mathbb{C}^n)$ for $k > 1$.

Proof. (i) Suppose that $X^{(\ell)}$ is not a frame for $\mathrm{Sym}^\ell(V)$. Then $X^{(\ell)}$ does not span $\mathrm{Sym}^\ell(V)$ and there exists $g \in (\mathrm{span}(X^{(\ell)}))^\perp \subset \mathrm{Sym}^\ell(V)$. Take some $h \in \mathrm{Sym}^{k-\ell}(V)$. Let $h = x_1^{\otimes(k-\ell)}$. Then

$$\begin{aligned} \left\langle g \otimes h, x_i^{\otimes k}\right\rangle &= \left\langle g \otimes h, x_i^{\otimes \ell} \otimes x_i^{\otimes(k-\ell)}\right\rangle \\ &= \left\langle g, x_i^{\otimes \ell}\right\rangle \left\langle h, x_i^{\otimes(k-\ell)}\right\rangle \\ &= 0 \cdot \left\langle x_1^{\otimes(k-\ell)}, x_i^{\otimes(k-\ell)}\right\rangle = 0 \end{aligned}$$

which is a contradiction since $g \otimes h \in \mathrm{Sym}^k(V)$ and $X^{(k)}$ is a frame for $\mathrm{Sym}^k(V)$ so that for any i, $\left\langle g \otimes h, x_i^{\otimes k}\right\rangle$ cannot be zero.

(ii) It has been shown in [7] that for almost every set of vectors $X = \{x_1, \ldots, x_m\}$ in $\mathbb{C}^n$, the rank of the grammian of $X^{(k)} = \{x_1^{\otimes k}, \ldots, x_m^{\otimes k}\}$ is $\binom{n+k-1}{k}$ when $m > \binom{n+k-1}{k}$. This means that the maximum number of linearly independent vectors in $X^{(k)}$ is $\binom{n+k-1}{k}$ which is the same as the dimension of $\mathrm{Sym}^k(\mathbb{C}^n)$ and hence $X^{(k)}$ is a frame for $\mathrm{Sym}^k(\mathbb{C}^n)$. □

3. Illustrative Examples

Example 1. Consider the space $V = \mathbb{C}^2$ over the field $\mathbb{C}$. Let $x_1 = [1, 0]^\mathsf{T}$, $x_2 = [0, 1]^\mathsf{T}$ and $x_3 = [1, 1]^\mathsf{T}$. The set $X = \{x_1, x_2, x_3\}$ is a frame for V with corresponding frame operator

$$F = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}.$$

The metric operator is

$$\mathcal{F} = F^*F = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

The eigenvalues of $\mathcal{F}$ are 1 and 3, which are the optimal lower and upper frame bounds respectively.

$$\mathcal{F}^{-1} = \frac{1}{3}\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

which is the metric operator of the dual frame The dual frame is denoted by $\tilde{X} = \{\tilde{x}_1, \tilde{x}_2, \tilde{x}_3\}$ where

$$\tilde{x}_1 = \mathcal{F}^{-1}x_1 = \left[\frac{2}{3}, -\frac{1}{3}\right]^\mathsf{T},$$

$$\tilde{x}_2 = \mathcal{F}^{-1}x_2 = \left[-\frac{1}{3}, \frac{2}{3}\right]^\mathsf{T}, \text{ and}$$

$$\tilde{x}_3 = \mathcal{F}^{-1}x_3 = \left[\frac{1}{3}, \frac{1}{3}\right]^\mathsf{T}.$$

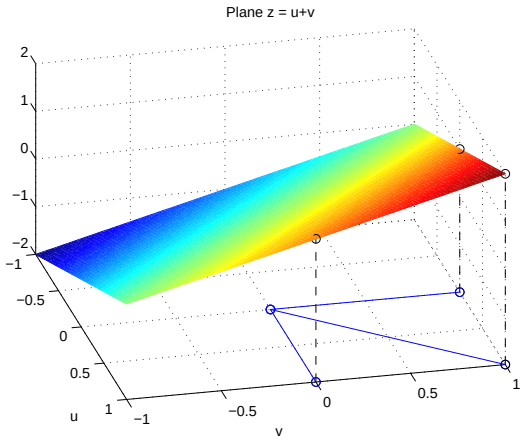


Figure 1: The plane $z = u+v$, a homogeneous polynomial of degree one.

Consider reconstruction of the homogeneous polynomial p of degree one in two variables defined by $p(u,v) = c^{(1)}u + c^{(2)}v$ from the three frame elements. Here $k = 1$, $n = 2$ and $m = 3$. Any $c = [c^{(1)}, c^{(2)}]^{\mathsf{T}} \in \mathbb{C}^2$ can be reconstructed via the frame reconstruction formula

$$c = \sum_{i=1}^{3} \langle x_i, c\rangle \, \tilde{x}_i. \tag{3}$$

Since $p(x_i) = \langle x_i, c\rangle$ and for this example $p(x_1) = c^{(1)}$, $p(x_2) = c^{(2)}$, and $p(x_3) = c^{(1)} + c^{(2)}$, the right side of (3) is

$$c^{(1)}\tilde{x}_1 + c^{(2)}\tilde{x}_2 + (c^{(1)} + c^{(2)})\tilde{x}_3 = [c^{(1)}, c^{(2)}]^{\mathsf{T}}.$$

This shows that the coefficients of $p(u,v)$ can be reconstructed from its samples at the frame elements. The polynomial $p(u,v) = u + v$ together with the sampling set is shown in Figure 1.

Example 2. If the homogeneous polynomial to be reconstructed is of degree two as given by $p(u,v) = c^{(1)}u^2 + c^{(2)}uv + c^{(3)}v^2$ then one considers the space $\mathrm{Sym}^2(\mathbb{C}^2) \subset \mathbb{C}^{\otimes 2}$. The dimension of $\mathrm{Sym}^2(\mathbb{C}^2)$ is three, which is the same as the dimension of $H_2(\mathbb{C}^2)$. Hence at least three sampling points are needed. Consider the same set of sampling points as in Example 1; i.e., $x_1 = [1,0]^{\mathsf{T}}$, $x_2 = [0,1]^{\mathsf{T}}$ and $x_3 = [1,1]^{\mathsf{T}}$. One can extend this set to $\mathbb{C}^{\otimes 2}$ by taking Kronecker products and restricting to $\mathrm{Sym}^2(\mathbb{C}^2)$ yields $x_1^{\otimes 2} = [1,0,0]^{\mathsf{T}}$, $x_2^{\otimes 2} = [0,0,1]^{\mathsf{T}}$, and $x_3^{\otimes 2} = [1,1,1]^{\mathsf{T}}$. Let $X^{(2)} = \{x_1^{\otimes 2}, x_2^{\otimes 2}, x_3^{\otimes 2}\}$. The polynomial p can be uniquely determined from its sample values at x_1, x_2 and x_3 because $c^{(1)} = p(x_1)$, $c^{(3)} = p(x_2)$, and $c^{(2)} = p(x_3) - p(x_2) - p(x_1)$. This means that $X^{(2)}$ is a frame for $\mathrm{Sym}^2(\mathbb{C}^2)$. The frame operator is

$$F = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

making the metric operator

$$\mathcal{F} = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 2 \end{bmatrix}.$$

The minimum and maximum eigenvalues of $\mathcal{F}$ are .2679 and 3.7321, which are the optimal lower and upper frame bounds respectively. The metric operator for the dual frame is

$$\mathcal{F}^{-1} = \begin{bmatrix} -1 & -1 & 0 \\ -1 & 3 & -1 \\ 0 & -1 & 1 \end{bmatrix}$$

making the dual frame $\widetilde{x_1^{\otimes 2}} = \mathcal{F}^{-1}x_1^{\otimes 2} = [1,-1,0]^{\mathsf{T}}$, $\widetilde{x_2^{\otimes 2}} = \mathcal{F}^{-1}x_2^{\otimes 2} = [0,-1,1]^{\mathsf{T}}$, and $\widetilde{x_3^{\otimes 2}} = \mathcal{F}^{-1}x_3^{\otimes 2} = [0,1,0]^{\mathsf{T}}$.
The polynomial $p(u,v) = c^{(1)}u^2 + c^{(2)}uv + c^{(3)}v^2$ satisfies $p(x_1) = c^{(1)}, p(x_2) = c^{(3)}$, and $p(x_3) = c^{(1)} + c^{(2)} + c^{(3)}$. The coefficients of p can be obtained from its samples at x_1, x_2, and x_3 by the frame reconstruction formula for $\mathrm{Sym}^2(\mathbb{C}^2)$; i.e.,

$$[c^{(1)}, c^{(2)}, c^{(3)}]^{\mathsf{T}} = p(x_1)\widetilde{x_1^{\otimes 2}} + p(x_2)\widetilde{x_2^{\otimes 2}} + p(x_3)\widetilde{x_3^{\otimes 2}}.$$

Example 3. Consider now the frame for $\mathbb{C}^2$ formed by $x_1 = [1,0]^{\mathsf{T}}$, $x_2 = [2,0]^{\mathsf{T}}$, and $x_3 = [0,1]^{\mathsf{T}}$. In this case, reconstruction of $p(u,v) = c^{(1)}u^2 + c^{(2)}uv + c^{(3)}v^2$ from samples $p(x_1)$, $p(x_2)$, and $p(x_3)$ is not generally possible, even though the number of samples is the same as the dimension of $H_2(\mathbb{C}^2)$. This is because x_1 and x_2 are scalar multiples of each other and the corresponding vectors in $\mathrm{Sym}^2(\mathbb{C}^2)$, $\{[1,0,0]^{\mathsf{T}}, [2,0,0]^{\mathsf{T}}, [0,0,1]^{\mathsf{T}}\}$ do not constitute a frame for $\mathrm{Sym}^2(\mathbb{C}^2)$. This is an example where the tensor powers of a frame for V do not frame $\mathrm{Sym}^k(V)$, even though the number of vectors is adequate.

Example 4. Reconstruction of homogeneous polynomials in $H_3(\mathbb{C}^2)$ requires at least four points, since the dimension of $\mathrm{Sym}^3(\mathbb{C}^2)$ and hence that of $H_3(\mathbb{C}^2)$ is four. Taking the frame $X = \{x_1, x_2, x_3, x_4\} = \{[1,0]^{\mathsf{T}}, [0,1]^{\mathsf{T}}, [1,1]^{\mathsf{T}}, [1,-1]^{\mathsf{T}}\}$ for $\mathbb{C}^2$, computing Kronecker products and restricting to $\mathrm{Sym}^3(\mathbb{C}^2)$ yields

$$\begin{aligned} X^{(3)} &= \{x_1^{\otimes 3}, x_2^{\otimes 3}, x_3^{\otimes 3}, x_4^{\otimes 3}\} \\ &= \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix} \right\}. \end{aligned}$$

A homogeneous polynomial of the form $p(u,v) = c^{(1)}u^3 + c^{(2)}u^2v + c^{(3)}uv^2 + c^{(4)}v^3$ can be reconstructed from its samples at these points as $c^{(1)} = p(1,0)$, $c^{(2)} = \frac{1}{2}(p(1,1) + p(1,-1) - 2p(1,0))$, $c^{(3)} = \frac{1}{2}(p(1,1) + p(1,-1) - 2p(1,0))$, and $c^{(4)} = p(0,1)$ so that $X^{(3)}$ constitutes a frame for $\mathrm{Sym}^3(\mathbb{C})^2$. The frame operator is

$$F = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix}$$

making the metric operator

$$\mathcal{F} = \begin{bmatrix} 3 & 0 & 2 & 0 \\ 0 & 2 & 0 & 2 \\ 2 & 0 & 2 & 0 \\ 0 & 2 & 0 & 3 \end{bmatrix}.$$

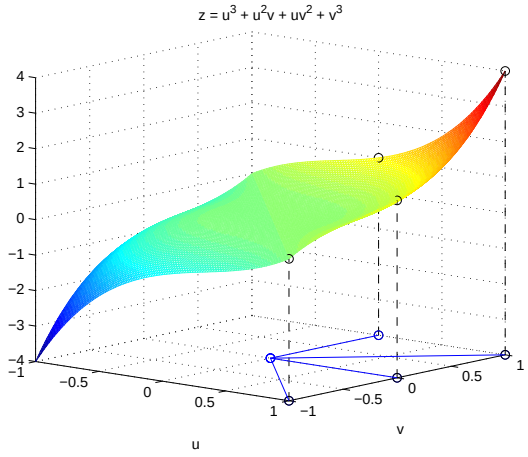


Figure 2: A homogeneous polynomial of degree three.

The optimal lower and upper frame bounds are $A = 0.4384$ and $B = 4.5616$. The metric operator of the dual frame is

$$\mathcal{F}^{-1} = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1.5 & 0 & -1 \\ -1 & 0 & 1.5 & 0 \\ 0 & -1 & 0 & -1 \end{bmatrix}.$$

The dual frame is $\widetilde{x_1^{\otimes 3}} = \mathcal{F}^{-1}x_1^{\otimes 3} = [1, 0, -1, 0]^\mathsf{T}$, $\widetilde{x_2^{\otimes 3}} = \mathcal{F}^{-1}x_2^{\otimes 3} = [0, -1, 0, 1]^\mathsf{T}$, $\widetilde{x_3^{\otimes 3}} = \mathcal{F}^{-1}x_3^{\otimes 3} = [0, .5, .5, 0]^\mathsf{T}$, and $\widetilde{x_4^{\otimes 3}} = \mathcal{F}^{-1}x_4^{\otimes 3} = [0, -.5, .5, 0]^\mathsf{T}$. The coefficients of a degree-three polynomial $p(u, v) = c^{(1)}u^3 + c^{(2)}u^2v + c^{(3)}uv^2 + c^{(4)}v^3$ are given by

$$[c^{(1)}, c^{(2)}, c^{(3)}, c^{(4)}]^\mathsf{T} = p(x_1)\widetilde{x_1^{\otimes 3}} + p(x_2)\widetilde{x_2^{\otimes 3}} + p(x_3)\widetilde{x_3^{\otimes 3}} + p(x_4)\widetilde{x_4^{\otimes 3}}.$$

Such a polynomial and the sampling points are shown in Figure 2.

Example 5. As the degree k or the dimension n gets larger numerical issues arise in calculating the inverse of the metric operator in order to get the dual frame that is needed for the reconstruction [4]. Ideally, one would like to construct tight frames for $\mathrm{Sym}^k(\mathbb{C}^n)$. Since the upper and lower frame bounds determine the numerical merits of a particular frame, it is interesting to observe how starting with a fixed frame for $\mathbb{C}^2$ the frame bounds change as this frame is extended to frames for $\mathrm{Sym}^k(\mathbb{C}^2)$ as k increases. Taking the frame for $\mathbb{C}^2$ to be $\{[1, 0]^\mathsf{T}, [0, 1]^\mathsf{T}, [1, 1]^\mathsf{T}, [1, -1]^\mathsf{T}\}$, the frame bounds for $\mathbb{C}^2$, $\mathrm{Sym}^2(\mathbb{C}^2)$, and $\mathrm{Sym}^3(\mathbb{C}^2)$ are tabulated below.

Space	Optimal lower frame bound A	Optimal upper frame bound B	B/A
$\mathbb{C}^2$	3	3	1
$\mathrm{Sym}^2(\mathbb{C}^2)$	1	5	5
$\mathrm{Sym}^3(\mathbb{C}^2)$	.4384	4.5616	10.4

In this particular case it appears that the ratio B/A increases as k, the degree of the polynomial increases.

4. Conclusions and Future Work

It has been shown that homogeneous polynomials of degree k in n variables can be reconstructed from their samples at elements of a frame for $\mathrm{Sym}^k(\mathbb{C}^n)$. Such a set can also be used to reconstruct n-variate homogeneous polynomials of all degrees ℓ where $1 \leqslant \ell < k$. In recent work [1], [9] conditions under which a smooth function can be approximated by homogeneous polynomials have been established. Combining these results to approximately reconstruct smooth functions from sampled data and a possible construction of tight frames for $\mathrm{Sym}^k(\mathbb{C}^n)$ will be given in a detailed version of this work. The metric operator and the grammian of a frame have the same non-zero eigenvalues. Also $\mathcal{G}^{\circ k}$, the grammian of $X^{(k)} = \{x_1^{\otimes k}, x_2^{\otimes k}, \ldots, x_m^{\otimes k}\}$, is the k-fold Hadamard product of $\mathcal{G}$, the grammian of $X = \{x_1, x_2, \ldots, x_m\}$. Relationship between the eigenvalues of $\mathcal{G}$ and $\mathcal{G}^{\circ k}$ ([2], [5], [6]) may be used to obtain information about the frame bounds for a frame for $\mathrm{Sym}^k(\mathbb{C}^n)$ which comes from a frame for $\mathbb{C}^n$, see Example 5.

5. Acknowledgments

The authors would like to thank John McDonald for useful discussions on the topic of this paper.

References:

[1] D. Benko and A. Kroó. A Weierstrass-type theorem for homogeneous polynomials. *Transactions of the American Mathematical Society*, 361(3):1645 – 1665, 2009.

[2] G. Cheng, X. Cheng, T. Huang, and T. Tam. Some bounds for the spectral radius of the Hadamard product of matrices. *Applied Mathematics E-Notes*, 5:202–209, 2005.

[3] S. Datta, S. D. Howard, and D. Cochran. Geometry of the Welch bounds. *IEEE Transactions on Information Theory*. In review.

[4] I. Daubechies. *Ten Lectures on Wavelets*. SIAM, 1992.

[5] M. Fang. Bounds on eigenvalues of the Hadamard product and the Fan product of matrices. *Linear Algebra and its Applications*, 425:7–15, 2007.

[6] E. I. Im. Narrower eigenbounds for Hadamard products. *Linear Algebra and its Applications*, 264:141–144, 1997.

[7] I. Peng and S. Waldron. Signed frames and Hadamard products of Gram matrices. *Linear Algebra and its Applications*, 347:131–157, 2002.

[8] R. Shaw. *Linear algebra and group representations*, volume 2. Academic Press, 1983.

[9] P. Varjú. Approximation by homogeneous polynomials. *Constructive Approximation*, 26:317 – 337, 2007.

[10] A. Zakhor and G. Alvstad. Two-dimensional polynomial interpolation from nonuniform samples. *IEEE Transactions on Signal Processing*, 40(1):169 – 180, 1992.

On sampling lattices with similarity scaling relationships

Steven Bergner [(1)], Dimitri Van De Ville[(2)], Thierry Blu[(3)], and Torsten Möller[(1)]

(1) GrUVi-Lab, Simon Fraser University, Burnaby, Canada.
(2) BIG, Ecole Polytechnique Fédérale de Lausanne, Switzerland.
(3) The Chinese University of Hong Kong, Hong Kong, China.
sbergner@cs.sfu.ca, thierry.blu@m4x.org, Dimitri.VanDeVille@epfl.ch, torsten@cs.sfu.ca

Abstract:

We provide a method for constructing regular sampling lattices in arbitrary dimensions together with an integer dilation matrix. Subsampling using this dilation matrix leads to a similarity-transformed version of the lattice with a chosen density reduction. These lattices are interesting candidates for multidimensional wavelet constructions with a limited number of subbands.

1. Primer on sampling lattices and related work

A sampling lattice is a set of points $\{\mathbf{Rk} : \mathbf{k} \in \mathbb{Z}^n\} \subset \mathbb{R}^n$ that is closed under addition and inversion. The non-singular generating matrix $\mathbf{R} \in \mathbb{R}^{n\times n}$ contains basis vectors in its columns. Lattice points are uniquely indexed by $\mathbf{k} \in \mathbb{Z}^n$ and the neighbourhood around each sampling point is identical. This makes them suitable sampling patterns for the reconstruction in shift-invariant spaces.
Subsampling schemes for lattices are expressed in terms of a dilation matrix $\mathbf{K} \in \mathbb{Z}^{n\times n}$ forming a new lattice with generating matrix $\mathbf{RK}$. The reduction rate in sampling density corresponds to

$$|\det \mathbf{K}| = \alpha^n = \delta \in \mathbb{Z}^+. \tag{1}$$

Dyadic subsampling discards every second sample along each of the n dimensions resulting in a $\delta = 2^n$ reduction rate. To allow for fine-grained scale progression we are particularly interested in low subsampling rates, such as $\delta = 2$ or 3.
As discussed by van de Ville et al. [8], the 2D quincunx subsampling is an interesting case permitting a two-channel relation. With the implicit assumption of only considering subsets of the Cartesian lattice it is shown that a similarity two-channel dilation may not extend for $n > 2$.
Here, we show that by permitting more general basis vectors in $\mathbb{R}^n$ the desired fixed-rate dilation becomes possible for any n. Our construction produces a variety of lattices making it possible to include additional quality criteria into the search as they may be computed from the Voronoi cell of the lattice [9] including packing density and expected quadratic quantization error (second order moment). Agrell et al. [1] improve efficiency for the computation by extracting Voronoi relevant neighbours. Another possible sampling quality criterion appears in the work of Lu et al. [4] in form of an analytic alias-free sampling condition that is employed in a lattice search.

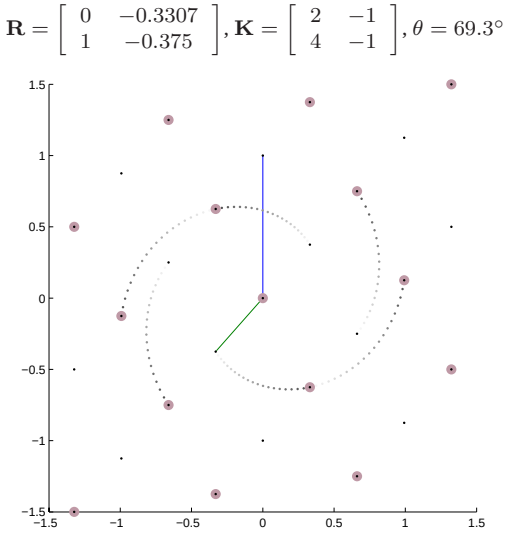


Figure 1: 2D lattice with basis vectors and subsampling as given by $\mathbf{R}$ and $\mathbf{K}$ in the diagram title. The spiral shaped points correspond to a sequence of fractional subsamplings $\mathbf{RK}^s$ for $s = 0..1$ with the notable feature that for $s = 1$ one obtains a subset of the original lattice sites shown as thick dots. This repeats for any further integer power of $\mathbf{K}$, each time reducing the sample density by $|\det \mathbf{K}| = 2$.

2. Lattice construction

We are looking for a non-singular lattice generating matrix $\mathbf{R}$ that, when sub-sampled by a dilation matrix $\mathbf{K}$ with reduction rate $\delta = \alpha^n$, results in a similarity-transformed version of the same lattice, that is, it can be scaled and rotated by a matrix $\mathbf{Q}$ with $\mathbf{Q}^T\mathbf{Q} = \alpha^2\mathbf{I}$. An illustration of a subsampling resulting in a rotation by $\theta = \arccos \frac{1}{2\sqrt{2}}$ in 2D is given in Figure 1. Formally, this kind of relationship can be expressed as

$$\mathbf{QR} = \mathbf{RK} \tag{2}$$

leading to the observation that subsampling $\mathbf{K}$ and scaled rotation $\mathbf{Q}$ are related by a similarity transform

$$\mathbf{R}^{-1}\mathbf{QR} = \mathbf{K}. \tag{3}$$

Using a matrix $\mathbf{J}_2 = \begin{bmatrix} 1 & j \\ 1 & -j \end{bmatrix}$ it is possible to diagonalize a 2D rotation matrix by the following similarity transform

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = \mathbf{J}_2^{-1} \begin{bmatrix} e^{j\theta} & 0 \\ 0 & e^{-j\theta} \end{bmatrix} \mathbf{J}_2 = \mathbf{J}_2^{-1} \mathbf{\Delta} \mathbf{J}_2. \tag{4}$$

Using this observation to replace the scaled rotation matrix $\mathbf{Q}$ in Equation 3 leads to

$$\begin{aligned} \mathbf{K} &= \mathbf{R}^{-1}\mathbf{Q}\mathbf{R} \\ \mathbf{K} &= \alpha \mathbf{R}^{-1}\mathbf{J}_n^{-1}\mathbf{S}\mathbf{\Delta}\mathbf{S}^{-1}\mathbf{J}_n\mathbf{R} \\ \mathbf{K} &= \alpha \mathbf{P}\mathbf{\Delta}\mathbf{P}^{-1} \end{aligned} \tag{5}$$

with

$$\begin{aligned} \mathbf{R} &= \mathbf{J}_n^{-1}\mathbf{S}\mathbf{P}^{-1} \\ \mathbf{Q} &= \alpha \mathbf{J}_n^{-1}\mathbf{\Delta}\mathbf{J}_n. \end{aligned} \tag{6}$$

Thus, given a matrix $\mathbf{K}$ that has an eigen-decomposition corresponding to that of a uniformly scaled rotation matrix, we can compute the lattice generating matrix $\mathbf{R}$ as in Equation 6. The elements of the diagonal matrix $\mathbf{S}$ inserted in the construction of $\mathbf{R}$ scale the otherwise unit eigenvectors in the columns of $\mathbf{P}$. Below, we will refer to this construction as function formRQ($\mathbf{K}, \mathbf{S}$) using $\mathbf{S} = \mathbf{I}$ by default.

2.1 Constructing suitable dilation matrices $\mathbf{K}$

The eigenvalues of $\mathbf{K}, \mathbf{\Delta}$ and $\mathbf{Q}$ impose restrictions on their shared characteristic polynomial $d(\lambda) = \det(\mathbf{K} - \lambda\mathbf{I}) = \sum_{k=0}^{n} c_k \lambda^k$ as discussed in the appendix. For the case $n = even$ with the only non-zero integer coefficients $c_0 = \delta,\ c_{n/2}^2 < 4\delta,\ c_n = 1$ this leaves a finite number of different options for $c_{n/2}$. The case $n = odd$ permits a single possible polynomial with non-zero coefficients $c_0 = -\delta,\ c_n = 1$. For these monic polynomials it is possible to directly construct a candidate $\mathbf{K}$ via the companion matrix ([6], p. 192)

$$\mathbf{K} = \begin{bmatrix} 0 & & & & -c_0 \\ 1 & 0 & & & -c_1 \\ & 1 & 0 & & \vdots \\ & & \ddots & \ddots & -c_{n-2} \\ & & & 1 & -c_{n-1} \end{bmatrix}. \tag{7}$$

This allows to construct a lattice fulfilling the self-similar subsampling condition for any dimensionality n, one for every possible characteristic polynomial.

With this starting point it is possible to construct additional suitable dilation matrices via a similarity transform with a unimodular matrix $\mathbf{T}$

$$\mathbf{K}_T = \mathbf{T}\mathbf{K}\mathbf{T}^{-1} = \mathbf{P}_T\mathbf{\Delta}\mathbf{P}_T^{-1}. \tag{8}$$

Using a unimodular rather than any non-singular $\mathbf{T}$ guarantees that $\mathbf{T}^{-1}$ is also unimodular following from the fact that $\mathbf{T}^{-1}$ can be constructed from the adjugate (the transposed co-factor matrix) of $\mathbf{T}$. Thus, $\mathbf{K}_T$ remains an integer matrix by this transform. Possible generators for this unimodular group are discussed in ([5], pp. 23). Our implementation, referred to as function genUnimodular(n), uses a construction of $\mathbf{T} = \mathbf{L}\mathbf{U}$ from several random integer lower and upper triangular matrices having ones on their diagonal.

It is not guaranteed that all possible $\mathbf{K}$ for a given characteristic polynomial can be generated through a similarity transform with some $\mathbf{T}$. However, formRQ($\mathbf{K}_T$) provides numerous non-equivalent $\mathbf{R}_T$ lattice generators. Among them it is possible to apply further criteria to select the "best" lattice.

An alternative to transforming $\mathbf{K}$ is the eigenvector scaling by diagonal matrix $\mathbf{S}$ in Equation 6. Using non-unit scaling allows to produce further lattices for any given $\mathbf{K}$ resulting in an n-dimensional continuous search space.

2.2 Construction Algorithm

The steps for constructing lattices with the desired subsampling matrices are summarized in algorithm 1.

The function compoly(n, α, C) is defined in the appendix. A possible implementation for the function genUnimodular(n) is described in Section 2.1 and formRQ($\mathbf{K}$) is defined below Equation 6.

Algorithm 1 genLattices(n, δ)

```
Llist ← {}
Ks ← genKompans(n, δ)
Ts ← genUnimodular(n) ∪{I}
for all K ∈ Ks do
    for all T ∈ Ts do
        K_T = TKT^-1
        (R_T, Q_T) ← formRQ(K_T)
        Llist← Llist∪{(K_T, R_T, Q_T)}
    end for
end for
return Llist
```

Algorithm 2 genKompans(n, δ)

```
Ks = {}
if n is even then
    for all C ∈ ℤ : C^2 < 4δ do
        Ks ← Ks ∪ compoly(n, δ^(1/n), C)
    end for
else {n is odd}
    Ks ← {compoly(n, δ^(1/n))}
end if
return Ks
```

It should be noted that the list of lattices returned by genLattices may contain several equivalent copies of the same lattice. A Gram matrix implicitly represents angles between basis vectors as $\mathbf{A} = \mathbf{R}^T\mathbf{R}$. Two lattices $\mathbf{R}_1$ and $\mathbf{R}_2$, scaled to have the same determinant, are equivalent if their Gram matrices are related via $\mathbf{A}_1 = \mathbf{T}^T\mathbf{A}_2\mathbf{T}$ with a unimodular matrix $\mathbf{T} \in \mathbb{Z}^{n\times n}$ and $|\det \mathbf{T}| = 1$. Determining this unimodular matrix is known to be a difficult problem, as it for instance also occurs when relating the adjacency matrices of two supposedly isomorphic graphs. Hence, our current method employs a simpler necessary test for equivalence by comparing the first few elements

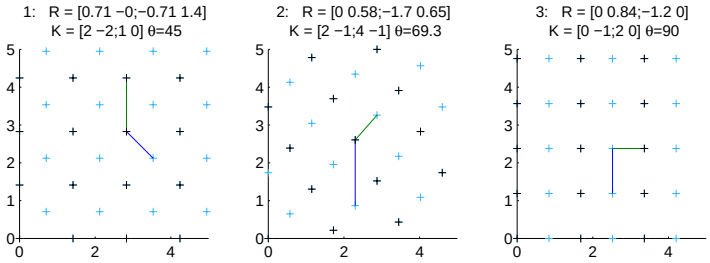


Figure 2: Three non-equivalent 2D lattices obtained for a design with dilation matrices having $|\det \mathbf{K}| = 2$. The lattice on the left is the well known quincunx sampling with a $\theta = 45^\circ$ rotation. The other two are new schemes with different rotation angles. The black markers show the sample positions that are retained after subsampling by $\mathbf{K}$.

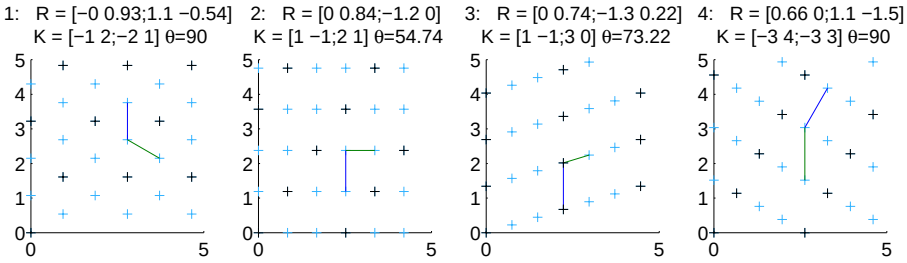


Figure 3: Three non-equivalent 2D lattices obtained for a design with dilation matrices having $|\det \mathbf{K}| = 3$. The lattice on the left is the well known hexagonal lattice with a $\theta = 30^\circ$ rotation. The other three are new schemes with different rotation angles.

of the set $q(\mathbf{A}) = \{\mathbf{k}^T\mathbf{A}\mathbf{k} : \mathbf{k} \in \mathbb{Z}^n\}$ using the Gram matrices of the respective lattices. If the sorted lists $q(\mathbf{A}_1)$ and $q(\mathbf{A}_2)$ disagree in any element, $\mathbf{R}_1$ and $\mathbf{R}_2$ are not equivalent ([5], p. 60). It is possible to restrict the set of indices $\mathbf{k} \in \mathbb{Z}^n$ to the Voronoi relevant neighbours [1]. Further, since these neighbours determine the hyperplanes bounding the Voronoi polytope of the lattice, they can also be used for a sufficient test for equivalence.

3. Constructions for different dimensions and subsampling ratios

For the 2D case we have created lattices permitting a reduction rate 2 in Figure 2 and rate 3 in Figure 3. In both cases, familiar examples arise in the quincunx and the hex lattice for the respective ratios.

A search of 3D lattices enjoying the self-similar subsampling property with rate 2 dilations resulted in 53 non-equivalent cases. These lattices were compared in terms of their dimensionless second order moments, corresponding to the expected squared vector quantization error ([2], p. 451). When performing the continuous optimization mentioned at the end of Section 2.1, all of these cases converged to the same optimum lattice shown in Figure 4. The dimensionless second order moment for the Voronoi Cell of this lattice is $G = 0.081904$. For comparison, the Cartesian cube has $G_{cc} = 0.0833$ and the truncated octahedron of the BCC lattice has $G_{bcc} = 0.0785$.

4. Discussion and potential applications

The current formation of candidate matrices $\mathbf{K}$ based on similarity transforms of one valid example is not guaranteed to produce all possible solutions. For 2D and 3D we also employed an exhaustive search over a range of integer matrices with values in $[-3, 3]$ resulting in the same number of non-equivalent 2D cases as the construction via $\mathbf{K}_T$. However, for dimensionality $n > 3$ the exhaustive search had to be replaced by a random sampling of integer matrices ultimately rendering the method infeasible for $n > 5$. In that light the current construction via scaled eigenvectors of the companion matrix is a significant improvement as it allows to produce a large number of non-equivalent lattices for any dimensionality.

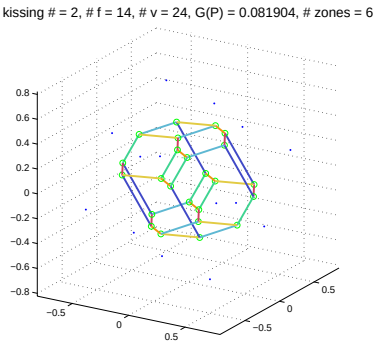


Figure 4: The best 3D lattice obtained for a design with dilation matrices having $|\det \mathbf{K}| = 2$. The letters f and v in the title line indicate faces and vertices, respectively.

Our subsampling schemes may have applications for multidimensional wavelet transforms [7]. Another direction for possible investigation is the construction of sparse grids that are employed in the context of high-dimensional integration and approximation adapting to smoothness conditions of the underlying function space [3].

Appendix: Characteristic polynomial of a scaled rotation matrix in $\mathbb{R}^n$

The similarity relationship between $\mathbf{K}$ and $\mathbf{Q}$ in Equation 2 implies that they share the same characteristic polynomial $d(\lambda) = \det(\mathbf{K} - \lambda\mathbf{I}) = \det(\mathbf{Q} - \lambda\mathbf{I})$ leading to an agreement in eigenvalues $d(\lambda_k) = 0$ and determinant $d(0)$ ([6], p. 184). Further, since $\mathbf{K}$ is an integer matrix the polynomial $d(\lambda) \in \mathbb{Z}[\lambda]$ has integer coefficients c_k.
In order to find integer matrices $\mathbf{K}$ with the eigenvalues of a scaled rotation matrix, it will be important to distinguish the two different forms of the diagonal matrix $\mathbf{\Delta}$ in Equation 4 and 5 for the case $n = \textit{even}$

$$\mathbf{\Delta} = \text{diag}[e^{j\theta_1}\ e^{-j\theta_1}\ \ldots e^{j\theta_{n/2}}\ e^{-j\theta_{n/2}}]$$

and the case $n = \textit{odd}$

$$\mathbf{\Delta} = \text{diag}[1\ e^{j\theta_1}\ e^{-j\theta_1} \ldots e^{j\theta_{(n-1)/2}}\ e^{-j\theta_{(n-1)/2}}]$$

with analogue block-wise constructions for $\mathbf{J}_n$.
For dimensionality $n = even$ the characteristic polynomial of $\mathbf{K}$ and $\mathbf{Q}$ fulfills

$$\begin{aligned} d(\lambda) &= \prod_{k=1}^{n/2} (\alpha e^{j\theta_k} - \lambda)(\alpha e^{-j\theta_k} - \lambda) \\ &= \prod_{k=1}^{n/2} (\alpha^2 - 2\lambda\alpha\cos\theta_k + \lambda^2) \\ &= \prod_{k=1}^{n/2} \left[(\frac{\alpha^4}{\lambda^2} - 2\frac{\alpha^3}{\lambda}\cos\theta_k + \alpha^2)\frac{\lambda^2}{\alpha^2} \right] \\ &= d\left(\frac{\alpha^2}{\lambda}\right)\left(\frac{\lambda}{\alpha}\right)^n \end{aligned} \tag{9}$$

Thus, if

$$\begin{aligned} d(\lambda) &= \sum_{k=0}^{n} c_k \lambda^k \\ &= \sum_{k=0}^{n} c_k \left(\frac{\alpha^2}{\lambda}\right)^k \left(\frac{\lambda}{\alpha}\right)^n \\ &= \sum_{k=0}^{n} c_{n-k} \alpha^{n-2k} \lambda^k \\ &\Leftrightarrow c_k = \alpha^{n-2k} c_{n-k} = \delta^{1-\frac{2k}{n}} c_{n-k}. \end{aligned} \tag{10}$$

If $c_k \neq 0$ and $c_k, \delta \in \mathbb{Z}$ then $\delta^{1-\frac{2k}{n}} \in \mathbb{Q}$. This is impossible for $0 < 2k < n$, assuming small values of δ, such as $2, 3$ or any simple product of primes. This implies that $c_k = c_{n-k} = 0$ for $k = 1, 2, \ldots \frac{n}{2} - 1$. For $k = \frac{n}{2}$ the c_k can be non-zero leading to

$$d(\lambda) = \lambda^n + C\lambda^{\frac{n}{2}} + \alpha^n \tag{11}$$

with the requirement that $C^2 < 4\alpha^n$ so that the complex eigenvalues $d(\lambda_k) = 0$ are evenly distributed on the complex circle of radius $|\lambda_k| = \alpha$.
For dimensionality $n = \textit{odd}$ the polynomial fulfills

$$\begin{aligned} d(\lambda) &= (\alpha - \lambda) \prod_{k=1}^{(n-1)/2} (\alpha e^{j\theta_k} - \lambda)(\alpha e^{-j\theta_k} - \lambda) \\ &\Rightarrow d(\lambda) = -\left(\frac{\lambda}{\alpha}\right)^n d\left(\frac{\alpha^2}{\lambda}\right) \end{aligned} \tag{12}$$

Thus, if

$$\begin{aligned} d(\lambda) &= \sum_{k=0}^{n} c_k \lambda^k \\ &= -\sum_{k=0}^{n} c_k \left(\frac{\alpha^2}{\lambda}\right)^k \left(\frac{\lambda}{\alpha}\right)^n \\ &= -\sum_{k=0}^{n} c_{n-k} \alpha^{n-2k} \lambda^k \\ &\Leftrightarrow c_k = -\alpha^{n-2k} c_{n-k} = -\delta^{1-\frac{2k}{n}} c_{n-k}. \end{aligned} \tag{13}$$

By the same reasoning as for the even case, $c_k = 0$ for all $k = 1, 2, \ldots \frac{n-1}{2}$ resulting in only one possible characteristic polynomial

$$d(\lambda) = \lambda^n - \alpha^n. \tag{14}$$

To refer to the above procedure we will invoke a function compoly(n, α, C) that returns a companion matrix (Equation 7) with a characteristic polynomial as in Equation 11 or 14.

References:

[1] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger. Closest point search in lattices. *Information Theory, IEEE Transactions on*, 48(8):2201–2214, August 2002.

[2] J.H. Conway and N.J.A. Sloane. *Sphere Packings, Lattices and Groups. – 3rd ed*. Springer, 1999.

[3] M. Griebel. Sparse grids and related approximation schemes for higher dimensional problems. In L. Pardo, A. Pinkus, E. Suli, and M.J. Todd, editors, *Foundations of Computational Mathematics (FoCM05), Santander*, pages 106–161. Cambridge University Press, 2006.

[4] Y.M. Lu, M.N. Do, and R.S. Laugesen. A Computable Fourier Condition Generating Alias-Free Sampling Lattices. *IEEE Transactions on Signal Processing*, 57(5):(15 pages), May 2009.

[5] M. Newman. *Integral Matrices*. Academic Press, 1972. See `http://www.dleex.com/read/?3907` for a digital copy.

[6] L.N. Trefethen and D. Bau III. *Numerical Linear Algebra*. SIAM, 1997.

[7] D. Van De Ville, T. Blu, and M. Unser. Isotropic polyharmonic B-Splines: Scaling functions and wavelets. *IEEE Transactions on Image Processing*, 14(11):1798–1813, November 2005.

[8] D. Van De Ville, T. Blu, and M. Unser. On the multidimensional extension of the quincunx subsampling matrix. *IEEE Signal Processing Letters*, 12(2):112–115, February 2005.

[9] E. Viterbo and E. Biglieri. Computing the Voronoi cell of a lattice: The diamond-cutting algorithm. *Information Theory, IEEE Trans. on*, 42(1):161–171, 1996.

General Perturbations of Sparse Signals in Compressed Sensing

Matthew A. Herman and Thomas Strohmer

Department of Mathematics, University of California, Davis, CA 95616-8633, USA.
{mattyh,strohmer}@math.ucdavis.edu

Abstract:

We analyze the Basis Pursuit recovery of signals when observing sparse data with general perturbations. Previous studies have only considered partially perturbed observations $\boldsymbol{A}\boldsymbol{x}+\boldsymbol{e}$. Here, $\boldsymbol{x}$ is a K-sparse signal which we wish to recover, $\boldsymbol{A}$ is a measurement matrix with more columns than rows, and $\boldsymbol{e}$ is simple *additive* noise. Our model also incorporates perturbations $\boldsymbol{E}$ (which result in *multiplicative* noise) to the matrix $\boldsymbol{A}$ in the form of $(\boldsymbol{A}+\boldsymbol{E})\boldsymbol{x}+\boldsymbol{e}$. This completely perturbed framework extends the previous work of Candès, Romberg and Tao on stable signal recovery from incomplete and inaccurate measurements. Our results show that, under suitable conditions, the stability of the recovered signal is limited by the noise level in the observation. Moreover, this accuracy is within a constant multiple of the best-case reconstruction using the technique of least squares.

1. Introduction

Employing the techniques of compressed sensing (CS) to recover signals with a sparse representation has enjoyed a great deal of attention over the last 5–10 years. The initial studies considered an ideal unperturbed scenario:

$$\boldsymbol{b} \;=\; \boldsymbol{A}\boldsymbol{x}. \tag{1}$$

Here $\boldsymbol{b} \in \mathbb{C}^m$ is the observation vector, $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ ($m \leq n$) is a full-rank measurement matrix or system model, and $\boldsymbol{x} \in \mathbb{C}^n$ is the signal of interest which has a K-sparse representation (i.e., it has no more than K nonzero coefficients) under some fixed basis. More recently researchers have included an *additive* noise term $\boldsymbol{e}$ into the received signal [1, 2, 4, 8], creating a *partially perturbed model*:

$$\hat{\boldsymbol{b}} \;=\; \boldsymbol{A}\boldsymbol{x}+\boldsymbol{e} \tag{2}$$

This type of noise generally models simple, uncorrelated errors in the data or at the receiver/sensor.

As far as we can tell, practically no research has been done yet on perturbations $\boldsymbol{E}$ to the matrix $\boldsymbol{A}$. Our ***completely perturbed model*** extends (2) by incorporating a perturbed sensing matrix in the form of

$$\hat{\boldsymbol{A}} = \boldsymbol{A}+\boldsymbol{E}.$$

It is important to consider this kind of noise since it can account for precision errors when applications call for physically implementing the matrix $\boldsymbol{A}$ in a sensor. When $\boldsymbol{A}$ represents a system model, such as in the context of radar [7] or telecommunications, then $\boldsymbol{E}$ can absorb errors in assumptions made about the transmission channel, as well as quantization errors arising from the discretization of analog signals. In general, these perturbations can be characterized as *multiplicative noise*, and are more difficult to analyze than simple additive noise since they are correlated with the signal of interest. To see this, simply substitute $\boldsymbol{A} = \hat{\boldsymbol{A}} - \boldsymbol{E}$ in (2); there will be an extra noise term $\boldsymbol{E}\boldsymbol{x}$. (Note that it makes no difference whether we account for the perturbation $\boldsymbol{E}$ on the "encoding side" (2), or on the "decoding side" (7). The model used here was chosen so as to agree with the conventions of classical perturbation theory which we use in Section 4.)

1.1 Assumptions and Notation

Without loss of generality, assume the original data $\boldsymbol{x}$ to be a K-sparse vector for some fixed K. Denote $\sigma^{(K)}_{\max}(\boldsymbol{Y})$, $\|\boldsymbol{Y}\|^{(K)}_2$, and $\text{rank}^{(K)}(\boldsymbol{Y})$ respectively as the maximum singular value, spectral norm, and rank over all K-column submatrices of a matrix $\boldsymbol{Y}$. Similarly, $\sigma^{(K)}_{\min}(\boldsymbol{Y})$ is the minimum singular value over all K-column submatrices of $\boldsymbol{Y}$. Let the perturbations in (2) be relatively bounded by

$$\frac{\|\boldsymbol{E}\|^{(K)}_2}{\|\boldsymbol{A}\|^{(K)}_2} \leq \varepsilon^{(K)}_{\boldsymbol{A}}, \quad \frac{\|\boldsymbol{e}\|_2}{\|\boldsymbol{b}\|_2} \leq \varepsilon_{\boldsymbol{b}} \tag{3}$$

with $\|\boldsymbol{A}\|^{(K)}_2, \|\boldsymbol{b}\|_2 \neq 0$. In the real world we are only interested in the case where both $\varepsilon^{(K)}_{\boldsymbol{A}}, \varepsilon_{\boldsymbol{b}} < 1$.

2. CS ℓ_1 Perturbation Analysis

2.1 Previous Work

In the *partially perturbed scenario* (i.e., $\boldsymbol{E} = \boldsymbol{0}$ in (2)) we are concerned with solving the *Basis Pursuit* (BP) problem [3]:

$$\boldsymbol{z}^\star = \arg\min_{\hat{\boldsymbol{z}}} \|\hat{\boldsymbol{z}}\|_1 \;\text{ s.t. }\; \|\boldsymbol{A}\hat{\boldsymbol{z}} - \hat{\boldsymbol{b}}\|_2 \leq \varepsilon' \tag{4}$$

for some $\varepsilon' \geq 0$.

The *restricted isometry property* (RIP) [2] for any matrix $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ defines, for each integer $K = 1, 2, \ldots,$

the *restricted isometry constant* (RIC) δ_K, which is the smallest nonnegative number such that

$$(1-\delta_K)\|\boldsymbol{x}\|_2^2 \;\le\; \|\boldsymbol{A}\boldsymbol{x}\|_2^2 \;\le\; (1+\delta_K)\|\boldsymbol{x}\|_2^2 \quad (5)$$

holds for any K-sparse vector $\boldsymbol{x}$. In the context of the RIC, we observe that $\|\boldsymbol{A}\|_2^{(K)} = \sigma_{\max}^{(K)}(\boldsymbol{A}) = \sqrt{1+\delta_K}$, and $\sigma_{\min}^{(K)}(\boldsymbol{A}) = \sqrt{1-\delta_K}$.

Assuming K-sparse $\boldsymbol{x}$, $\delta_{2K} < \sqrt{2}-1$ and $\|\boldsymbol{e}\|_2 \le \varepsilon'$, Candès has shown in Theorem 1.2 of [1] that the solution to (4) obeys

$$\|\boldsymbol{z}^\star - \boldsymbol{x}\|_2 \;\le\; C_{\text{BP}}\,\varepsilon' \quad (6)$$

for some constant C_{BP}.

2.2 Incorporating nontrivial perturbation $\boldsymbol{E}$

Now assume the ***completely perturbed*** situation with $\boldsymbol{E}, \boldsymbol{e} \neq \boldsymbol{0}$ in (2). In this case the BP problem of (4) can be generalized to include a different decoding matrix $\hat{\boldsymbol{A}}$:

$$\boldsymbol{z}^\star = \arg\min_{\hat{\boldsymbol{z}}} \|\hat{\boldsymbol{z}}\|_1 \quad \text{s.t.} \quad \|\hat{\boldsymbol{A}}\hat{\boldsymbol{z}} - \hat{\boldsymbol{b}}\|_2 \;\le\; \varepsilon'_{\boldsymbol{A},K,\boldsymbol{b}} \quad (7)$$

for some $\varepsilon'_{\boldsymbol{A},K,\boldsymbol{b}} \ge 0$. The following two theorems summarize our results.

Theorem 1 (RIP for $\hat{\boldsymbol{A}}$)**.** *For any $K = 1, 2, \ldots,$ assume and fix the RIC δ_K associated with $\boldsymbol{A}$, and the relative perturbation $\varepsilon_{\boldsymbol{A}}^{(K)}$ associated with $\boldsymbol{E}$ in (3). Then the RIC*

$$\hat{\delta}_K \;:=\; \big(1+\delta_K\big)\Big(1+\varepsilon_{\boldsymbol{A}}^{(K)}\Big)^2 - 1 \quad (8)$$

for matrix $\hat{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{E}$ is the smallest nonnegative constant such that

$$(1-\hat{\delta}_K)\|\boldsymbol{x}\|_2^2 \;\le\; \|\hat{\boldsymbol{A}}\boldsymbol{x}\|_2^2 \;\le\; (1+\hat{\delta}_K)\|\boldsymbol{x}\|_2^2 \quad (9)$$

holds for any K-sparse vector $\boldsymbol{x}$.

Remark 1. The flavor of the RIP is defined with respect to the square of the operator norm. That is, $(1-\delta_K)$ and $(1+\delta_K)$ are measures of the ***square*** of minimum and maximum singular values of $\boldsymbol{A}$, and similarly for $\hat{\boldsymbol{A}}$. In keeping with the convention of classical perturbation theory however, we defined $\varepsilon_{\boldsymbol{A}}^{(K)}$ in (3) just in terms of the operator norm (not its square). Therefore, the quadratic dependence of $\hat{\delta}_K$ on $\varepsilon_{\boldsymbol{A}}^{(K)}$ in (8) makes sense. Moreover, in discussing the spectrum of $\hat{\boldsymbol{A}}$, we see that it is really a *linear function* of $\varepsilon_{\boldsymbol{A}}^{(K)}$.

Theorem 2 (Completely perturbed observation)**.** *Fix the relative perturbations $\varepsilon_{\boldsymbol{A}}^{(K)}$, $\varepsilon_{\boldsymbol{A}}^{(2K)}$ and $\varepsilon_{\boldsymbol{b}}$ in (3). Assume that the RIC for matrix $\boldsymbol{A}$ satisfies $\delta_{2K} < \sqrt{2}\big(1+\varepsilon_{\boldsymbol{A}}^{(2K)}\big)^{-2} - 1$. Set*

$$\varepsilon'_{\boldsymbol{A},K,\boldsymbol{b}} := \Big(c\varepsilon_{\boldsymbol{A}}^{(K)} + \varepsilon_{\boldsymbol{b}}\Big)\,\|\boldsymbol{b}\|_2, \quad (10)$$

where $c = \frac{\sqrt{1+\delta_K}}{\sqrt{1-\delta_K}}$. If $\boldsymbol{x}$ is K-sparse, then the solution to the BP problem (7) obeys

$$\|\boldsymbol{z}^\star - \boldsymbol{x}\|_2 \;\le\; C_{BP}\,\varepsilon'_{\boldsymbol{A},K,\boldsymbol{b}}, \quad (11)$$

where

$$C_{BP} := \frac{4\sqrt{1+\delta_{2K}}\,\Big(1+\varepsilon_{\boldsymbol{A}}^{(2K)}\Big)}{1-(\sqrt{2}+1)\left(\big(1+\delta_{2K}\big)\Big(1+\varepsilon_{\boldsymbol{A}}^{(2K)}\Big)^2 - 1\right)}. \quad (12)$$

Remark 2. Theorem 2 generalizes of Candès' results in [1] for K-sparse $\boldsymbol{x}$. Indeed, if matrix $\boldsymbol{A}$ is unperturbed, then $\boldsymbol{E} = \boldsymbol{0}$ and $\varepsilon_{\boldsymbol{A}}^{(K)} = 0$. It follows that $\hat{\delta}_K = \delta_K$ in (8), and the RIPs for $\boldsymbol{A}$ and $\hat{\boldsymbol{A}}$ coincide. Moreover, the condition in Theorem 2 reduces to $\delta_K < \sqrt{2}-1$, and the total perturbation (see (17)) collapses to $\|\boldsymbol{e}\|_2 \le \varepsilon'_{\boldsymbol{b}} := \varepsilon_{\boldsymbol{b}}\|\boldsymbol{b}\|_2$; both of these are identical to Candès' assumptions in (6). Finally, the constant C_{BP} in (12) reduces to the same as outlined in the proof of [1].

It is also interesting to examine the spectral effects due to the assumptions of Theorem 2. Namely, we want to be assured that the rank of submatrices of $\boldsymbol{A}$ are unaltered by the perturbation $\boldsymbol{E}$.

Lemma 1. *If the hypothesis of Theorem 2 is satisfied, then for any $k \le 2K$*

$$\sigma_{\max}^{(k)}(\boldsymbol{E}) \;<\; \sigma_{\min}^{(k)}(\boldsymbol{A}), \quad (13)$$

and therefore

$$\mathit{rank}^{(k)}(\hat{\boldsymbol{A}}) \;=\; \mathit{rank}^{(k)}(\boldsymbol{A}).$$

This fact is necessary (although, not explicitly stated) in the least squares analysis Section 4.

The utility of Theorems 1 and 2 can be understood with two simple numerical examples. Suppose that measurement matrix $\boldsymbol{A}$ in (2) is designed to have an RIC of $\delta_{2K} = 0.100$. Assume, however, that its physical implementation will experience a worst-case relative error of $\varepsilon_{\boldsymbol{A}}^{(2K)} = 5\%$. Then from (8) we can design a matrix $\hat{\boldsymbol{A}}$ with RIC $\hat{\delta}_{2K} = 0.213$ to be used in (7) which will yield a solution whose accuracy is guaranteed by (11) with $C_{\text{BP}} = 9.057$. Note from (12), we see that if there had been no perturbation, then $C_{\text{BP}} = 5.530$.

Consider now a different example. Suppose instead that $\delta_{2K} = 0.200$ and $\varepsilon_{\boldsymbol{A}}^{(2K)} = 1\%$. Then $\hat{\delta}_{2K} = 0.224$ and $C_{\text{BP}} = 9.643$. Here, if $\boldsymbol{A}$ was unperturbed, then we would have had $C_{\text{BP}} = 8.473$.

These numerical examples show how the stability constant C_{BP} of the BP solution gets worse with perturbations to $\boldsymbol{A}$. It must be stressed however, that they represent worst-case instances. It is well-known in the CS community that better performance is normally achieved in practice.

2.3 Numerical Simulations

Numerical simulations were conducted as follows. Gaussian matrices of size 128×512 were randomly generated in MATLAB. The entries of matrix $\boldsymbol{A}$ were normally distributed $\mathcal{N}(0, \sigma_A^2)$ where $\sigma_A^2 = 1/128$, while those of matrix $\boldsymbol{E}$ were $\mathcal{N}(0, \sigma_E^2)$ with $\sigma_E^2 = \varepsilon_{\boldsymbol{A}}^2/128$. The parameter $\varepsilon_{\boldsymbol{A}}$ is a measure of the relative perturbation of matrix $\boldsymbol{A}$ and took on values $\{0, 0.01, 0.05, 0.10\}$. Next, a random

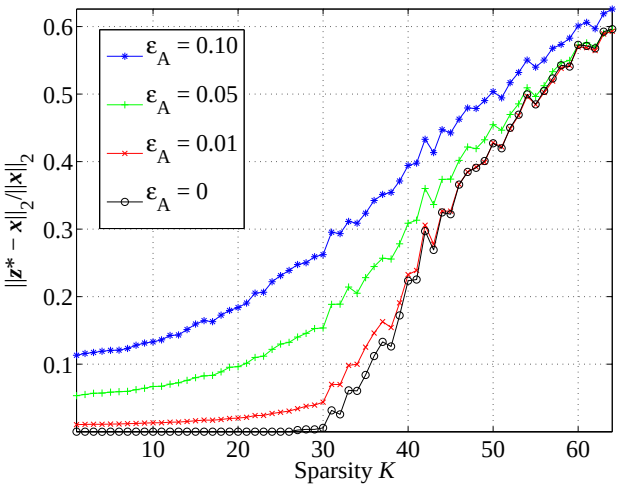


Figure 1: Average (100 trials) relative error of BP solution $\boldsymbol{z}^\star$ with respect to K-sparse $\boldsymbol{x}$ vs. Sparsity K for different relative perturbations $\varepsilon_{\boldsymbol{A}}$ of $\boldsymbol{A} \in \mathbb{C}^{128\times 512}$ (and $\varepsilon_{\boldsymbol{b}} = 0$) .

vector $\boldsymbol{x}$ of sparsity $K = 1, \ldots, 64$ was randomly generated (nonzero entries uniformly distributed with $\mathcal{N}(0,1)$) and $\hat{\boldsymbol{b}} = \boldsymbol{A}\boldsymbol{x}$ in (2) was created (note, we set $\boldsymbol{e} = \boldsymbol{0}$ so as to focus on the effect of perturbation $\boldsymbol{E}$). Given $\hat{\boldsymbol{b}}$ and $\hat{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{E}$, the BP program (7) was implemented with cvx software [5]. For each value of $\varepsilon_{\boldsymbol{A}}$ and K, 100 trials were performed.

Fig. 1 shows the average relative error $\|\boldsymbol{z}^\star - \boldsymbol{x}\|_2/\|\boldsymbol{x}\|_2$ as a function of K for each $\varepsilon_{\boldsymbol{A}}$. As a reference, the ideal, noise-free case can be seen for $\varepsilon_{\boldsymbol{A}} = 0$. It is interesting to notice that all perturbations, including $\varepsilon_{\boldsymbol{A}} = 0$, experience significant jumps simultaneously at several places, such as $K = 31, 42, 43, 44$, etc. Now fix a particular value of $K \leq 30$ and compare the relative error for the three nonzero values of $\varepsilon_{\boldsymbol{A}}$. It is clear that the error scales roughly linearly with $\varepsilon_{\boldsymbol{A}}$. This empirical study essentially confirms the conclusion of Theorem 2, that the stability of the BP solution scales linearly with $\varepsilon_{\boldsymbol{A}}^{(K)}$ (i.e., the singular values of $\boldsymbol{E}$).

Note that better performance in theory and in simulation can be achieved if BP is used solely to determine the support of the solution. Then we can use least squares to find a better result. This is similar to the the best-case, oracle least squares solution discussed in Section 4.

3. Proofs

3.1 Proof Sketch of Theorem 1

From the triangle inequality, (5) and (3) we have

$$\begin{aligned} \|\hat{\boldsymbol{A}}\boldsymbol{x}\|_2^2 &\leq \left(\|\boldsymbol{A}\boldsymbol{x}\|_2 + \|\boldsymbol{E}\boldsymbol{x}\|_2\right)^2 && (14)\\ &\leq \left(\sqrt{1+\delta_K} + \|\boldsymbol{E}\|_2^{(K)}\right)^2 \|\boldsymbol{x}\|_2^2 && (15)\\ &\leq (1+\delta_K)\left(1+\varepsilon_{\boldsymbol{A}}^{(K)}\right)^2 \|\boldsymbol{x}\|_2^2. && (16)\end{aligned}$$

Moreover, this inequality is sharp for the following reasons:

- Equality occurs in (14) if $\boldsymbol{E}$ is a multiple of $\boldsymbol{A}$.
- Equality occurs in (15) whenever $\boldsymbol{x}$ is in the direction of the vector associated with the value $(1+\delta_K)$ in the RIP for $\boldsymbol{A}$.
- Equality occurs in (16) since, in this hypothetical case, we assume that $\boldsymbol{E} = \beta\boldsymbol{A}$ for some $0 < \beta < 1$. Therefore, the relative perturbation $\varepsilon_{\boldsymbol{A}}^{(K)}$ in (3) no longer represents a worst-case deviation (i.e., the ratio $\frac{\|\boldsymbol{E}\|_2^{(K)}}{\|\boldsymbol{A}\|_2^{(K)}} = \beta =: \varepsilon_{\boldsymbol{A}}^{(K)}$).

The full details of this proof can be found in [6] □

3.2 Bounding the perturbed observation

Before proceeding, we need some sense of the size of the total perturbation incurred by $\boldsymbol{E}$ and $\boldsymbol{e}$. We don't know *a priori* the exact values of $\boldsymbol{E}$, $\boldsymbol{x}$, or $\boldsymbol{e}$. But we can find an upper bound in terms of the relative perturbations in (3). The main goal in the following lemma is to remove the total perturbation's dependence on the input $\boldsymbol{x}$.

Lemma 2 (Total perturbation bound)**.** *Set* $\varepsilon'_{\boldsymbol{A},K,\boldsymbol{b}} := \left(c\varepsilon_{\boldsymbol{A}}^{(K)} + \varepsilon_{\boldsymbol{b}}\right)\|\boldsymbol{b}\|_2$, *where* $c = \frac{\sqrt{1+\delta_K}}{\sqrt{1-\delta_K}}$, *and* $\varepsilon_{\boldsymbol{A}}^{(K)}$ *and* $\varepsilon_{\boldsymbol{b}}$ *are defined in (3). Then the total perturbation obeys*

$$\|\boldsymbol{E}\boldsymbol{x}\|_2 + \|\boldsymbol{e}\|_2 \leq \varepsilon'_{\boldsymbol{A},K,\boldsymbol{b}} \qquad (17)$$

for all K*-sparse* $\boldsymbol{x}$.

Proof. From (1), (5) and (3) we have

$$\begin{aligned} \|\boldsymbol{E}\boldsymbol{x}\|_2 + \|\boldsymbol{e}\|_2 &= \left(\frac{\|\boldsymbol{E}\boldsymbol{x}\|_2}{\|\boldsymbol{A}\boldsymbol{x}\|_2} + \frac{\|\boldsymbol{e}\|_2}{\|\boldsymbol{b}\|_2}\right)\|\boldsymbol{b}\|_2 \\ &\leq \left(\frac{\|\boldsymbol{E}\|_2^{(K)}\|\boldsymbol{x}\|_2}{\sqrt{1-\delta_K}\,\|\boldsymbol{x}\|_2} + \frac{\|\boldsymbol{e}\|_2}{\|\boldsymbol{b}\|_2}\right)\|\boldsymbol{b}\|_2 \\ &\leq \left(c\varepsilon_{\boldsymbol{A}}^{(K)} + \varepsilon_{\boldsymbol{b}}\right)\|\boldsymbol{b}\|_2 \end{aligned}$$

for all $\boldsymbol{x}$ which are K-sparse. □

Note that the results in this paper can easily be expressed in terms of the perturbed observation by replacing

$$\|\boldsymbol{b}\|_2 \leq \frac{\|\hat{\boldsymbol{b}}\|_2}{1-\varepsilon_{\boldsymbol{b}}}.$$

This can be useful in practice since one normally only has access to $\hat{\boldsymbol{b}}$.

3.3 Proof Sketch of Theorem 2

We duplicate the techniques used in Candès' proof of Theorem 1.2 in [1], but with decoding matrix $\boldsymbol{A}$ replaced by $\hat{\boldsymbol{A}}$. Set the BP minimizer in (7) as $\boldsymbol{z}^\star = \boldsymbol{x} + \boldsymbol{h}$. Here, $\boldsymbol{h}$ is the perturbation from the true solution $\boldsymbol{x}$ induced by $\boldsymbol{E}$ and $\boldsymbol{e}$. Instead of Candès' (9), we determine that the image of $\boldsymbol{h}$ under $\hat{\boldsymbol{A}}$ is bounded by

$$\begin{aligned} \|\hat{\boldsymbol{A}}\boldsymbol{h}\|_2 &\leq \|\hat{\boldsymbol{A}}\boldsymbol{z}^\star - \hat{\boldsymbol{b}}\|_2 + \|\hat{\boldsymbol{A}}\boldsymbol{x} - \hat{\boldsymbol{b}}\|_2 \\ &\leq 2\,\varepsilon'_{\boldsymbol{A},K,\boldsymbol{b}} \end{aligned}$$

which follows from the BP constraint in (7) as well as $\boldsymbol{x}$ being a feasible solution (i.e., it satisfies Lemma 2). The rest of this proof can be found in [6] □

3.4 Proof of Lemma 1

Assume the hypothesis of Theorem 2. It is easy to show that this implies

$$\|\boldsymbol{E}\|_2^{(2K)} \;<\; \sqrt[4]{2} \;-\; \sqrt{1+\delta_{2K}}.$$

Simple algebraic manipulation then confirms that

$$\sqrt[4]{2} \;-\; \sqrt{1+\delta_{2K}} \;<\; \sqrt{1-\delta_{2K}} \;=\; \sigma_{\min}^{(2K)}(\boldsymbol{A}).$$

Therefore, (13) holds with $k = 2K$. Further, for any $k \leq 2K$ we have $\sigma_{\max}^{(k)}(\boldsymbol{E}) \leq \sigma_{\max}^{(2K)}(\boldsymbol{E})$ and $\sigma_{\min}^{(2K)}(\boldsymbol{A}) \leq \sigma_{\min}^{(k)}(\boldsymbol{A})$, which proves the lemma. □

4. Classical ℓ_2 Perturbation Analysis

Let the subset $T \subseteq \{1, \ldots, n\}$ have cardinality $|T| = K$, and note the following *T-restrictions*: $\boldsymbol{A}_T \in \mathbb{C}^{m\times K}$ denotes the submatrix consisting of the columns of $\boldsymbol{A}$ indexed by the elements of T, and similarly for $\boldsymbol{x}_T \in \mathbb{C}^K$.

Suppose the "oracle" case where we already know the support T of K-sparse $\boldsymbol{x}$. By assumption, we are only interested in the case where $K \leq m$ in which $\boldsymbol{A}_T$ has full rank. Given the completely perturbed observation of (2), the least squares problem consists of solving:

$$\boldsymbol{z}_T^{\#} \;=\; \operatorname*{argmin}_{\hat{\boldsymbol{z}}_T} \|\hat{\boldsymbol{A}}_T\hat{\boldsymbol{z}}_T - \hat{\boldsymbol{b}}\|_2.$$

Since we know the support T, it is trivial to extend $\boldsymbol{z}_T^{\#}$ to $\boldsymbol{z}^{\#} \in \mathbb{C}^n$ by zero-padding on the complement of T. Our goal is to see how the perturbations $\boldsymbol{E}$ and $\boldsymbol{e}$ affect $\boldsymbol{z}^{\#}$.

More discussion on the oracle least squares analysis can be found in [6]. In the end, we find using the same $\varepsilon'_{\boldsymbol{A},K,\boldsymbol{b}}$ in (10) that its stability is

$$\|\boldsymbol{z}^{\#} - \boldsymbol{x}\|_2 \;\leq\; C_{\text{LS}}\,\varepsilon'_{\boldsymbol{A},K,\boldsymbol{b}} \tag{18}$$

where $C_{\text{LS}} := 1/\sqrt{1-\delta_K}$.

4.1 Comparison of LS with BP

Now, we can compare the accuracy of the least squares solution in (18) with the accuracy of the BP solution found in (11). In both cases the error bound is of the form

$$C\varepsilon'_{\boldsymbol{A},K,\boldsymbol{b}}.$$

A detailed numerical comparison of C_{LS} with C_{BP} is not entirely valid, nor illuminating. This is due to the fact that we assumed the oracle setup in the the least squares analysis, which is the best that one could hope for. In this sense, the least squares solution we examined here can be considered a "best, worst-case" scenario. In contrast, the BP solution really should be thought of as a "worst, of the worst-case" scenarios.

The important thing to glean is that the accuracy of the BP solution, like the least squares solution, is on the order of the noise level $\varepsilon'_{\boldsymbol{A},K,\boldsymbol{b}}$ in the perturbed observation. This is an important finding since, in general, no other recovery algorithm can do better than the oracle least squares solution. These results are analogous to the comparison by Candès, Romberg and Tao in [2], although they only consider the case of additive noise $\boldsymbol{e}$.

5. Conclusion

We introduced a general perturbed model for CS, and found the conditions under which BP could stably recover the original data. This completely perturbed model extends previous work by including a multiplicative noise term in addition to the usual additive noise term. We only considered K-sparse signals, however these results can be extended to also include compressible signals (see [6]).

Simple numerical examples were given which demonstrated how the multiplicative noise reduced the accuracy of the recovered BP solution. In terms of the spectrum of the perturbed matrix $\hat{\boldsymbol{A}}$, we showed that the penalty on $\hat{\delta}_K$ was a graceful, linear function of the relative perturbation $\varepsilon_{\boldsymbol{A}}^{(K)}$. Numerical simulations were performed with $\varepsilon_{\boldsymbol{b}} = 0$ and appear to confirm the conclusion of Theorem 2, that the BP solution scales linearly with $\varepsilon_{\boldsymbol{A}}^{(K)}$.

We also found that the rank of $\hat{\boldsymbol{A}}$ did not exceed the rank of $\boldsymbol{A}$ under the assumed conditions. This permitted an analysis of the oracle least squares solution which showed that its accuracy, like the BP solution, was limited by the total noise in the observation.

Acknowledgment

This work was partially supported by NSF Grant No. DMS-0811169 and NSF VIGRE Grant No. DMS-0636297.

References:

[1] E. J. Candès. The restricted isometry property and its implications for compressed sensing. *Académie des Sciences*, I(346):589–592, 2008.

[2] E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59:1207–1223, 2006.

[3] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal Sci. Comput.*, 20(1):33–61, 1999.

[4] D. L. Donoho, M. Elad, and V. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inf. Theory*, 52(1):6–18, Jan. 2006.

[5] M. Grant, S. Boyd, and Y. Ye. `cvx`: Matlab software for disciplined convex programming. `http://www.stanford.edu/~boyd/cvx/`.

[6] M. A. Herman and T. Strohmer. General Deviants: An analysis of perturbations in compressed sensing. `http://www.math.ucdavis.edu/~mattyh/publications.html`.

[7] M. A. Herman and T. Strohmer. High-resolution radar via compressed sensing. *To appear in IEEE Trans. Signal Processing*, Jun. 2009.

[8] J. A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inf. Theory*, 51(3):1030–1051, Mar. 2006.

Analysis of High-Dimensional Signal Data by Manifold Learning and Convolutions

Mijail Guillemard [(1)] and Armin Iske [(1)]

(1) Department of Mathematics, University of Hamburg, D-20146 Hamburg, Germany.
guillemard@math.uni-hamburg.de, iske@math.uni-hamburg.de

Abstract:

A novel concept for the analysis of high-dimensional signal data is proposed. To this end, customized techniques from manifold learning are combined with convolution transforms, being based on wavelets. The utility of the resulting method is supported by numerical examples concerning low-dimensional parameterizations of scale modulated signals and solutions to the wave equation at varying initial conditions.

1. Introduction

Recent advances in nonlinear dimensionality reduction and manifold learning have provided new methods for the analysis of high-dimensional signals. In this problem, a very large data set $U \subset \mathbb{R}^n$ of scattered points is given, where the data points are assumed to lie on a compact submanifold $\mathcal{M}$ of $\mathbb{R}^n$, i.e. $U \subset \mathcal{M} \subset \mathbb{R}^n$. Moreover, the dimension $k = \dim(\mathcal{M})$ of $\mathcal{M}$ is assumed to be much smaller than the dimension of the ambient space $\mathbb{R}^n$, $k \ll n$. Now, the primary goal in the dimensionality reduction is the construction of a low-dimensional representation of the data U.

In this paper, a novel concept for signal data analysis through dimensionality reduction is proposed. To this end, suitable techniques from manifold learning are combined with convolution transforms. Moreover, another important ingredient is a (suitable) projection map $P : \mathbb{R}^n \to \mathbb{R}^k$ that finally outputs the desired low-dimensional representation for U. Note that for the sake of approximation quality, we need to preserve intrinsic geometrical and topological properties of the manifold $\mathcal{M}$, and so the construction of the composite dimensionality reduction method requires particular care. In the proposed data analysis, the geometric distortion of the manifold, being incurred by the chosen convolution transform, plays a key role.

We remark that similar concepts from differential geometry are enjoying increasing interest in related applications of sampling theory, including surface reconstruction in reverse engineering and image analysis [5]. Further related concepts can be found in classical dimensionality reduction schemes, such as in *principal component analysis* and *multidimensional scaling*, while more recent techniques are including *Isomap* and *LLE methods* [4, 7] *Local Tangent Space Alignment* (LTSA) [6], *Sample Logmaps* [1], and, most recently, *Riemannian Normal Coordinates* [2, 3].

The outline of the paper is as follows. In the following Section 2, the main ingredients of the proposed nonlinear dimensionality reduction scheme, especially the construction of the convolution and projection map, are explained. Then, in Section 3 relevant aspects concerning distortion analysis are addressed. Finally, Section 4 shows the good performance of the resulting nonlinear dimensionality reduction method. To this end, numerical examples concerning low-dimensional parameterization of scale modulated signals and solutions to the wave equation at varying initial conditions are illustrated.

2. Construction of the Data Analysis

Given a set of signals $U = \{u_i\}_{i=1}^m \subset \mathcal{M}$, that we assume to lie in (or near) a low-dimensional Riemannian compact submanifold $\mathcal{M}$, of $\mathbb{R}^n$, we wish to analyse the given data for the purpose of dimensionality reduction. Therefore, we assume that there is an embedding $A : \Omega \to \mathcal{M}$, giving a parameterization of $\mathcal{M}$, where the domain $\Omega \subset \mathbb{R}^d$ lies in a low-dimensional Euclidean space $\mathbb{R}^d$, i.e., $d \ll n$. But the parameter domain Ω is unknown. Therefore, the goal of dimensionality reduction is to find a sufficiently accurate approximation Ω' of Ω, through which the desired low-dimensional representation for U is obtained.

We remark that the construction of the data analysis is required to depend on intrinsic geometrical and topological properties of the manifold $\mathcal{M}$. To this end, we apply a particular convolution transform $T : \mathcal{M} \to \mathcal{M}_T$, $\mathcal{M}_T = \{T(p) \ : \ p \in \mathcal{M}\}$, to each of the data sites u_i, followed by a suitable projection $P : \mathcal{M}_T \to \Omega'$, yielding a nonlinear data transformation for dimensionality reduction. The following diagram reflects our concept.

$$\begin{array}{ccc} \Omega \subset \mathbb{R}^d & \xrightarrow{A} & U \subset \mathcal{M} \subset \mathbb{R}^n \\ & & \Big\downarrow T \\ \Omega' \subset \mathbb{R}^d & \xleftarrow[P]{} & U_T \subset \mathcal{M}_T \subset \mathbb{R}^n \end{array} \tag{1}$$

Note that both the construction of the transformation T and the projection need particular care. Indeed, in order to maintain the intrinsic geometrical properties of the manifold $\mathcal{M}$, it is required to investigate the curvature distortion of $\mathcal{M}$ under the transform T. For this purpose, convolution filters are powerful tools for the construction of

suitable signal transforms T. This is supported by our numerical results in Section 4., where wavelet transforms are used for a customized construction of T.
Finally, let us remark that standard methods in signal processing rely on on special characteristics of a discrete-time signal $u_k \in \mathbb{R}^n$, such as frequency content, time duration, phase and amplitude information, etc. In typical application scenarios, signal data are not just isolated items of information, but they are rather incorporating correlations reflecting characteristic properties of the sampled object. Therefore, when designing customized signal transforms, one should exploit available context information on characteristic properties of the target object in order to improve the quality of the data analysis. In our particular application scenario, special emphasis needs to be placed on intrinsic geometrical properties of the manifold $\mathcal{M}$, where a preprocessing distortion analysis of the curvature is of vital importance.

3. Curvature Distortion Analysis

Our main objective is to estimate the curvature distortion in the geometry of the manifold $\mathcal{M}$ incurred by the application of the linear transformation $T : \mathcal{M} \to \mathcal{M}_T$, where T may, for instance, representing a wavelet or a convolution filter. To this end, we first need to evaluate relevant effects on the geometrical deformation of $\mathcal{M}$ under various specific transformations T. This then amounts to constructing suitable transformations T which are well-adapted to the characteristic properties of the specific data. Preferable choices for $T : \mathcal{M} \to \mathcal{M}_T$ are diffeomorphisms, in which case $\dim(\mathcal{M}) = \dim(\mathcal{M}_T)$.

3.1 Sectional Curvature Distortions

In general, a fundamental invariant of a manifold with respect to its isometries are the sectional curvatures. This concept is derived from the idea of the Gaussian curvature in the setting of 2-manifolds, and is defined as

$$K_{\mathcal{M}} = \frac{< R(X,Y)Y, X >}{\|X\|^2\|Y\|^2 - < X, Y >^2},$$

for the *curvature tensor* R, defined for a triple of smooth vector fields X, Y, Z as

$$R(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z.$$

We recall that the affine connection (a Levi-Cevita connection for our situation) is a bilinear map

$$\nabla : C^\infty(\mathcal{M}, T\mathcal{M}) \times C^\infty(\mathcal{M}, T\mathcal{M}) \to C^\infty(\mathcal{M}, T\mathcal{M})$$

that can be expressed with the Christoffel symbols defined, for a particular system of local coordinates $(x_1, \ldots, x_n)$, as $\nabla_{\partial_i} \partial_j = \sum_{k=1}^n \Gamma_{ij}^k \partial_k$. The Christoffel symbols can be described with respect to the metric tensor via

$$\Gamma_{ij}^k = \frac{1}{2} \sum_{\ell=1}^m \left(\frac{\partial g_{j\ell}}{\partial x_i} + \frac{\partial g_{i\ell}}{\partial x_j} + \frac{\partial g_{ij}}{\partial x_\ell} \right) g^{\ell k}.$$

In order to estimate the distortion caused by the linear map $T : \mathcal{M} \to \mathcal{M}_T$, we compare the Gaussian curvatures between $\mathcal{M}$ and $\mathcal{M}_T$, denoted respectively $K_{\mathcal{M}}$, and $K_{\mathcal{M}_T}$,

$$D_K^T(p) = K_{\mathcal{M}}(p) - K_{\mathcal{M}_T}(T(p)) \quad \text{for } p \in \mathcal{M}.$$

If T is invertible, then the Gaussian curvature $K_{\mathcal{M}_T}$ in $\mathcal{M}_T$ can be computed as a function of the metric g in $\mathcal{M}$ by using a *pullback* of the curvature tensor R in $\mathcal{M}$ with respect to the inverse map $T^{-1} : \mathcal{M}_T \to \mathcal{M}$, or, equivalently, by using a *pushforward* of the curvature tensor R in $\mathcal{M}$ with respect to $T : \mathcal{M} \to \mathcal{M}_T$. An alternative strategy is to consider the composition of T with a particular system of local coordinates $(x_1, \ldots, x_n)$ of $\mathcal{M}$, along with the metric tensor

$$g_{ij}(p) = g_{ij}(x_1, \ldots, x_m) = \left\langle \frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j} \right\rangle.$$

When considering the linear transformation T representing the convolution filter, an important case is when T is represented by a Toeplitz matrix, with filter coefficients $H = (h_1, \ldots, h_m)$, i.e.,

$$T = \begin{bmatrix} h_1 & 0 & \ldots & 0 \\ h_2 & h_1 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ h_m & h_{m-1} & \ldots & h_1 \\ 0 & h_m & \ldots & h_2 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & h_m \end{bmatrix}.$$

Note that the curvature distortion caused by the map T will be controlled by the singular values of T, which due to the Toeplitz matrix structure, are obtained from the Fourier coefficients of H.
Now, our primary objective is to investigate the influence of the filter coefficients in H on the curvature distortion D_K^T. Moreover, we study filters being required to obtain a given curvature distortion. The latter is particularly useful for the adaptive construction of a low dimensional representation of U.

3.2 Curvature Distortions for Curves

As for the special case of a curve $r : I = [t_0, t_1] \to \mathbb{R}^m$, with arc-length parameterization $s(a,t) = \int_a^t \|r'(x)\|\, dx$, recall that the curvature of r is $k(s) = \|r''(s)\|$. For an arbitrary parameterizations of r, its curvature is given by

$$K^2 = \frac{\|\ddot{r}\|^2 \|\dot{r}\|^2 - < \ddot{r}, \dot{r} >^2}{(\|\dot{r}\|^2)^3}.$$

In the remainder of this section, we briefly discuss the curvature distortion under linear maps (e.g. convolution transform) and under smooth maps. To compute the curvature distortion of a curve $r : I = [t_0, t_1] \to \mathbb{R}^m$ under a linear map T, we consider the curvature of $r_T = \{Tr(t), t \in I\}$, computed as follows.

$$K_T^2 \equiv K_T^2(t) = \frac{\|T\ddot{r}\|^2 \|T\dot{r}\|^2 - < T\ddot{r}, T\dot{r} >^2}{(\|T\dot{r}\|^2)^3}. \quad (2)$$

As for the general case of smooth maps $F : \mathbb{R}^m \to \mathbb{R}^r$, the curvature distortion can be approximated by using the

Jacobian matrix J_F and its singular value decomposition,

$$\begin{aligned} J_F(p) &= \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(p) & \cdots & \frac{\partial f_1}{\partial x_m}(p) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_r}{\partial x_1}(p) & \cdots & \frac{\partial f_r}{\partial x_m}(p) \end{bmatrix} \\ &= U_F(p) D_F(p) V_F^T(p) \qquad \text{for } p \in \mathcal{M}. \end{aligned}$$

The curvature distortion of a curve $r : [t_0, t_1] \to \mathbb{R}^m$ under F can in this case be analyzed through the expression

$$K_F^2 \equiv K_F^2(p) = \frac{\|J_F\ddot{r}\|^2 \|J_F\dot{r}\|^2 - < J_F\ddot{r}, J_F\dot{r} >^2}{(\|J_F\dot{r}\|^2)^3},$$

where, unlike in the linear case (2), the Jacobian matrices J_F depend on $p \in \mathcal{M}$.

4. Numerical Examples

This section presents three different numerical examples to illustrate basic properties of the proposed analysis of high-dimensional signal data. Further details shall be discussed during the conference.

4.1 Low-dimensional parameterization of scale modulated signals

In this example, we illustrate the geometrical effect of a convolution transform for a set of functions lying on a curve embedded in a high dimensional space. More precisely, we analyze a scale modulated family of functions $U \subset \mathbb{R}^{64}$, parameterized by three values in $\Omega \subset \mathbb{R}^3$,

$$U = \left\{ f_{\alpha(t)} = \sum_{i=1}^{3} e^{-\alpha_i(t)(\cdot - b_i)^2} : \alpha(t) \in \Omega \right\}.$$

The parameter set for the scale modulation is given by the curve

$$\Omega = \left\{ \alpha(t) = (\alpha_1(t), \alpha_2(t), \alpha_3(t))^T \in \mathbb{R}^3, \ : t \in [t_0, t_1] \right\}.$$

Figure 1 (left) shows the parameter domain Ω, a star shaped curve in $\mathbb{R}^3$. A PCA projection in $\mathbb{R}^3$, applied to the set $U \subset \mathbb{R}^{64}$, is also displayed in Figure 1 (middle). The projection illustrates the curvature distortion caused by the nonlinear map $A : \Omega \subset \mathbb{R}^3 \to U \subset \mathbb{R}^{64}$, $A(\alpha(t)) = f_{\alpha(t)}$.

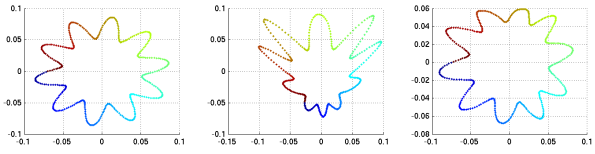


Figure 1: Parameter set $\Omega \subset \mathbb{R}^3$, data $U \subset \mathbb{R}^{64}$, and wavelet correction $T(U) \subset \mathbb{R}^{64}$.

Finally, Figure 1 (right), shows the resulting data transformation $T(U)$ using a Daubechies wavelet w.r.t. a specific band of the multiresolution analysis, resulting in a filtering process for each element in U. The resulting $T(U)$, presents a curvature correction that recovers the original geometry of Ω fairly well.
To explain the resulting curvature correction, we need to analyze the singular values and singular vectors of the convolution map T. In fact, the singular values of T can be viewed as scaling factors (stretching or shrinking) along corresponding axis in the (local) embedding of U. Moreover, the spectrum of T depends on the particular filter design.

4.2 Low dimensional parameterization of wave equation solutions

In this second example, we regard the one-dimensional wave equation

$$\frac{\partial u}{\partial t} = c^2 \frac{\partial u}{\partial x}, \quad 0 < x < 1, \ t \geq 0, \tag{3}$$

with initial conditions

$$u(0, x) = f(x), \ \frac{\partial u}{\partial t}(0, x) = g(x), \quad 0 \leq x \leq 1. \tag{4}$$

We make use of the previous example to construct a set of initial values (i.e. functions) parameterized by a star shaped curve $U_0 = U$. Our objective is to investigate the distortion caused by the evolution U_t of the solutions on given initial values U_0. Recall that the evolution of the wave equation is constituted by the set of solutions

$$\begin{aligned} U_t &= \{u_\alpha \equiv u_\alpha(t, x) : u_\alpha \text{ satisfying } (3) \text{ with} \\ &\qquad \text{initial condition } f \equiv f_\alpha \text{ in } (4) \text{ for } \alpha \in \Omega\}. \end{aligned}$$

Now, the solution of the wave equation can numerically be computed by using finite differences, yielding the iteration

$$u^{(j+1)} = Au^{(j)} + b^{(j)},$$

where for $\mu = \gamma \Delta t / (\Delta x)^2$, the iteration matrix is given by

$$A = \begin{bmatrix} 1-2\mu & \mu & & & & \\ \mu & 1-2\mu & \mu & & & \\ & \mu & 1-2\mu & \mu & & \\ & & \ddots & \ddots & \ddots & \\ & & & 0 & \mu & 1-2\mu \end{bmatrix}.$$

Recall that in the convergence analysis of the iteration, which can be rewritten as,

$$\begin{aligned} u^{(j+2)} &= Au^{(j+1)} + b^{(j+1)} \\ &= A(Au^{(j)} + b^{(j)}) + b^{(j+1)} \\ &= A^{(2)}u^{(j)} + Ab^{(j)} + b^{(j+1)}, \end{aligned}$$

the spectrum of the matrices A^k play a key role. In fact, due to the decomposition $A^k = UD^kU^T$, the geometrical distortion in the evolution of U_t depends on the evolution of the eigenvalues of A.

4.3 Topological Distortion via Filtering

In this final example, we illustrate one relevant phenomenon concerning the topological distortion caused by

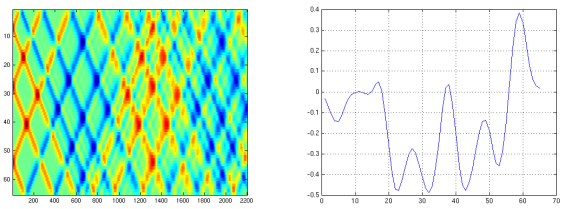


Figure 2: One solution of the wave equation $u(t,x)$ and one measurement $u(t_k,x)$, $t_k = 20$.

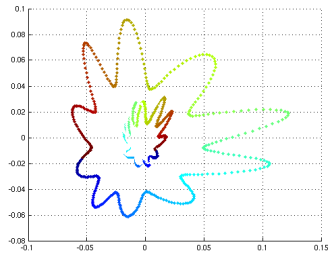


Figure 3: Curvature distortion of the initial manifold under the evolution of the wave equation. The outer curve represents the initial conditions U_0 while the inner curve reflects the corresponding solutions U_t for some time t.

the utilized convolution transformation. In this couple of two test cases, we take one 1-torus $\Omega_1 \subset \mathbb{R}^3$ and one 2-torus $\Omega_2 \subset \mathbb{R}^3$ as parameter space, respectively. As in the previous examples, we generate a corresponding set of scale modulation functions U_1 and U_2 (see Figure 4), using Ω_1 and Ω_2 as parameter domains. This gives, for $j = 1, 2$, two different data sets

$$U_j = \left\{ f_{\alpha^j(t)} = \sum_{i=1}^{3} e^{-\alpha_i^j(t)(\cdot - b_i^j)^2} : \alpha^j(t) \in \Omega_j \right\}.$$

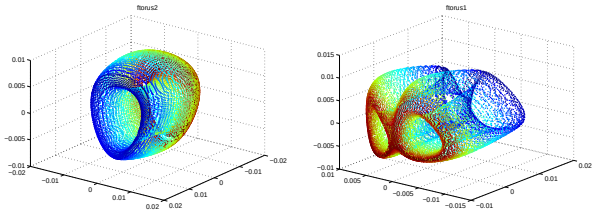


Figure 4: PCA projections of $U_1, U_2 \subset \mathbb{R}^{64}$ onto $\mathbb{R}^3$, generated by $\Omega_1, \Omega_2 \subset \mathbb{R}^3$, two tori of genus 1 and 2.

Now we combine the set U_1 and U_2 by

$$U = \left\{ f_t = f_{\alpha^1(t)} + f_{\alpha^2(t)} : \alpha^1(t) \in \Omega_1, \alpha^2(t) \in \Omega_2 \right\}.$$

The resulting projection of the data U is shown in Figure 5. For the purpose of illustration, we recover the sets U_1 and U_2 from U. Note that this is a rather challenging task, especially since the genus of surfaces U_1 and U_2 are different. Figure 6 shows the reconstructions of the two surfaces U_1 and U_2. Note that the both the geometrical and topological properties of U_1 and U_2 are recovered fairly well, which supports the good performance of our convolution transform yet once more. The reconstruction of the utilized convolution involves a selection of suitable bands from the corresponding wavelet multiresolution decomposition. Further details on this shall be explained during the conference.

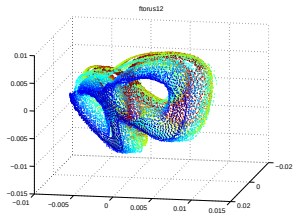


Figure 5: PCA projection of $U \subset \mathbb{R}^{64}$ onto $\mathbb{R}^3$.

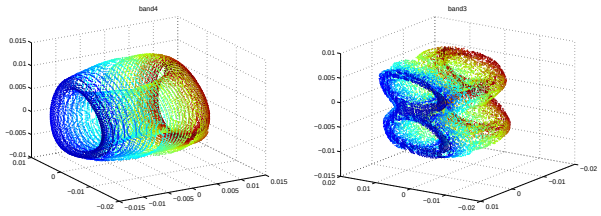


Figure 6: Reconstruction of U_1 (left), U_2 (right) from U.

5. Acknowledgments

The authors were supported by the priority program DFG-SPP 1324 of the *Deutsche Forschungsgemeinschaft.*

References:

[1] A. Brun, C. Westin, M. Herberthsson, and H. Knutsson. Sample logmaps: Intrinsic processing of empirical manifold data. *Proceedings of the (SSBA) Symposium on Image Analysis*, 1, 2006.

[2] A. Brun, C.-F. Westin, M. Herberthson, and H. Knutsson. Fast manifold learning based on riemannian normal coordinates. In *Proceedings of the SCIA;05*, pages 920–929, Joensuu, Finland, June 2005.

[3] T. Lin, H. Zha, and S.U. Lee. Riemannian Manifold Learning for Nonlinear Dimensionality Reduction. *Lecture Notes in Computer Science*, 3951:44, 2006.

[4] S.T. Roweis and L.K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding, 2000.

[5] E. Saucan, E. Appleboim, and Y.Y. Zeevi. Sampling and Reconstruction of Surfaces and Higher Dimensional Manifolds. *Journal of Mathematical Imaging and Vision*, 30(1):105–123, 2008.

[6] H. Zha and Z. Zhang. Principal Manifolds and Nonlinear Dimension Reduction via Local Tangent Space Alignment. *SIAM Journal of Scientific Computing*, 26(1):313–338, 2004.

[7] H. Zha and Z. Zhang. Continuum Isomap for manifold learnings. *Computational Statistics and Data Analysis*, 52(1):184–200, 2007.

Geometric Reproducing Kernels for Signal Reconstruction

Eli Appleboim [1], Emil Saucan [2] and Yehoshua Y. Zeevi [1]

(1) Technion, Dept. of Electrical Engineering
(2) Technion, Dept. of Mathematics
eliap@ee.technion.ac.il, semil@ee.technion.ac.il, zeevi@ee.technion.ac.il

Abstract:

In this paper we propose a smoothing method for non smooth signals, which control the geometry of a sampled signal. The signal is considered as a geometric object and the smoothing is done using a smoothing kernel function that controls the curvature of the obtained smooth signal in a close neighborhood of a metric curvature measure of the original signal.

1. Introduction

In [11], [12], a sampling scheme for signals that posses Riemannian geometric structure was introduced.It turns out that a variety of signals fall in this setting while gray scale images is just one such example. Rather then some Nyquist rate, the sampling scheme presented in [11], [12], is based on geometric characteristics of the sampled signals. Being precise, the following sampling theorem was proved.

Theorem 1 *Let $\Sigma^n, n \geq 2$ be a connected, not necessarily compact, smooth manifold, with finitely many compact boundary components. Then there exists a sampling scheme of Σ^n, with a proper density $\mathcal{D} = \mathcal{D}(p) = \mathcal{D}\left(\frac{1}{k(p)}\right)$, where $k(p) = \max\{|k_1|, ..., |k_{2n}|\}$, and where $k_1, ..., k_{2n}$ are the principal (normal) curvatures of Σ^n, at the point $p \in \Sigma^n$.*

While the assumed Riemannian structure relies on the assumption that the signal satisfies $\mathcal{C}^2$ smoothness criteria, the authors presented in [11], an extended version of Theorem 1 also for non smooth geometric signals, where the proposed strategy uses smoothing of the original signal. The following theorem was proved.

Theorem 2 *Let Σ be a connected, non-necessarily compact surface of class $\mathcal{C}^0$. Then, for any $\delta > 0$, there exists a δ-sampling of Σ, such that if $\Sigma_\delta \to \Sigma$, then $\mathcal{D}_\delta \to \mathcal{D}$, where $\mathcal{D}_\delta$ and $\mathcal{D}$ denote the densities of Σ_δ and Σ, respectively.*

In the above Theorem 2 Σ_δ is a smoothing of Σ obtained by a convolution of Σ with a partition of unity kernel. Such a kernel being very common for manifolds smoothing indeed guarantees that the resultant manifold is as smooth as we wish however, in this process we do not have any control on the curvature of the obtained manifold.
Some natural question raise in this context,

1. To what extent can we smooth the original signal, using such a reproducing kernel while assuming a predefined bounds on the curvature of the resultant manifold?

2. Can the reproducing kernel be made local, namely, can we have different kernel characteristics for different areas along the sampled signals, while being able to glue the smoothed signal along common boundaries?

3. In what way if at any, we can give affirmative answers to 1 and 2 that are adaptive to the signal? Meaning, how can we have good prior estimates for the desired curvature bounds?

This paper aims at answering the above questions. Note that answering question 1 is analogous to smoothen a signal to have a predefined frequency band-pass, using a band-pass filter as commonly done in signal processing for decades. Answering $1, 2, 3$ is equivalent to the use of filter banks with different band-pass characteristics. In all, giving affirmative answers to all above questions give rise to an adaptive non uniform sampling scheme for a variety of signals.
We will focus along the paper on signals that are do not admit a Riemannian structure but rather have a more general geometric structure of the so called Alexandrov spaces. We will term such signals as geometric-signals.

2. Preliminaries

In this section we will give some basic preliminary definitions and notations.

2.1 Alexandrov spaces

Definition 3 (Alexandrov - Toponogov) *[[9]] A complete metric space X, satisfies the **triangle comparison condition** w.r.t $\kappa \in \mathbb{R}$ if for every geodesic triangle $\Delta_{pqr} \in X$, there exists a **comparison triangle**, i.e. a triangle, $\Delta_{p'q'r'} \in \mathbb{M}^2_\kappa$, such that*

$$pq = p'q'; \;\; qr = q'r'; \;\; rp = r'p'$$

so that, for every point $s \in pr$ we have that

$$d_X(s, q) > d_{\mathbb{M}^2_\kappa}(s', q')$$

where $s' \in p'r'$ such that

$$ps = p's'; \;\; sr = s'r'$$

Where $\mathbb{M}^2_\kappa$ is a complete simply connected surface of ***constant curvature*** *κ.*

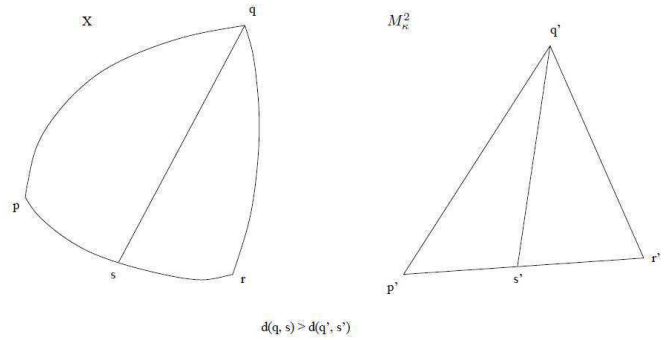


Figure 1: Comparison triangle.

Definition 4 *A complete metric space X, is an* ***Alexandrov space*** *of curvature $> \kappa$ iff*

1. *For all $x, y, \in X$ there exists a length minimizing curve γ joining x and y such that,*

   $$L(\gamma) = d_X(x, y);$$

   where L denotes the arc length of curves in X and d_X stands for the metric given on X. γ is called a ***minimal geodesic****.*

2. *X satisfies the triangle comparison condition for κ.*

3. $$dim_H X < \infty;$$

   dim_H = Hausdorff dimension.

Remark 5 *In a similar way, while reversing the direction of inequalities, one can define Alexandrov space of curvature $< \kappa$. For instance, in the comparison triangle condition, we will demand,*

$$d_X(s, q) < d_{\mathbb{M}^2_\kappa}(s', q')$$

Definition 6 (Gromov) *If X is an Alexandrov space of curvature $< \kappa$ and $\kappa \leq 0$ then X is called $CAT(\kappa)$-* ***space****. CAT = Cartan-Alexandrov-Toponogov.*

2.1.1 Examples:

1. Every complete Riemannian manifold of bounded sectional curvature.

2. The boundary of convex set in $\mathbb{R}^n$ is an Alexandrov space of curvature ≥ 0.

3. If X_i is a sequence of n-dimensional Alexandrov spaces of curv. $\geq \kappa$ then their Gromov-Hausdorff limit, if exists, is an Alexandrov space of curv. $\geq \kappa$ and dimension $\leq n$.

   If the limit of the above sequence is of dimension $< n$ we say the sequence **collapses**.

If X is an Alexandrov space then there exists a self-adjoint operator Δ, called the **Laplacian** defined on $\mathbb{L}^2(X)$ so that,

$$\int_X < \nabla u, \nabla v > d\mathcal{H}^n = \int_X v \nabla u d\mathcal{H}^n$$

where $\mathcal{H}^n$ is the n^{th} Hausdorff measure of X, $u \in \mathcal{D}(\Delta), v \in \mathcal{W}^{1,2}(X)$.

Theorem 7 ([6]) *1. If X is compact then the spectrum of Δ is discrete.*

2. There exists a continuous heat kernel $h_t(x, y)$ on X so that,

$$e^{-t\Delta}u(x) = \int_X h_t(x, y)u(y)d\mathcal{H}^n(y)$$

2.2 Approximations of manifolds

Let M be a complete Riemannian manifold of bounded sectional curvature. Let $p \in M$ be some point and let ϕ_i be some $\mathcal{C}^\infty$ kernel function supported on some ϵ_i-neighborhood of p. For example one can take ϕ to be partition of unity, heat kernel and others. Let M_i be the manifold obtained by convolution,

$$M_i = \int_M \phi_i * M d\mu;$$

Note that M_i is smooth in a δ_i neighborhood of p even if M fails to be smooth at p. Well known results (see for instance, [7]) in differential topology assert that,

$$\epsilon_i \to 0 \Rightarrow M_j \to M;$$

where convergence of manifolds is considered in the Gromov-Hausdorff topology. While the above result concerns the convergence on a topological level, in order to have curvature control we have to account for geometric convergence as well. This is guaranteed from the studies in [3], [4] and [10]. In [3], [4] it is proved that similar convergence to the above also exist for Betti numbers which are generalizations of Euler characteristic to all dimensions and are related to curvature through higher dimensional of Gauss-Bonnet type theorems [2]. In [10] the question of proper gluing of approximations in adjacent neighborhoods is addressed. It is shown that one can obtain geometric convergence in different neighborhoods V, U of the points p, q resp. so that, on the common boundary $\partial V \cap \partial U$ the approximations coincide. In addition, if we write the heat operator on a manifold, $\mathcal{N}$, as

$$e^{-t\Delta_\mathcal{N}} f(x),$$

where $f \in L^2(\mathcal{N})$ and $t > 0, x \in \mathcal{N}$, and $\Delta_\mathcal{N}$, denotes the Laplace-Beltrami operator associated with $\mathcal{N}$, then there is a smooth kernel function $K_\mathcal{N}$, such that,

$$e^{-t\Delta_\mathcal{N}} f(x) = \int_\mathcal{N} K_\mathcal{N}(t, x, y) f(y) dy;$$

In [3] convergence of the heat kernel is also achieved,

$$e^{-t\Delta_{M_i}} \to e^{-t\Delta_M}$$

3. Smoothing geometric signals with curvature control

In this section we present the results concerning questions $1, 2$ and 3 posed in the introduction. These results give us the ability to smoothen a geometric signal while having an adaptive control on obtained curvatures.

Definition 8 *We say that a signal is a* ***geometric signal*** *iff it admits a structure of an Alexandrov space for some* $\kappa \in \mathbb{R}$.

Let Σ be a geometric signal of sectional curvature bounded from below (above). Let $p \in \Sigma$ be a point, and $U(p) \subset \Sigma$ some compact neighborhood of p. Let

$$\kappa = \limsup K$$

such that $U(p)$ is an Alexandrov space of curvature $> K$.

3.1 Approximations of geometric signals

Theorem 9 ([1]) *Given a point p on Σ, there exists smooth local kernel ϕ_i as above, yielding a sequence of manifolds M_i, smooth inside an ϵ_i neighborhoods of p, such that*

1. $$M_i = \int_\Sigma \phi_i * \Sigma d\mu \to \Sigma,$$
   as $\epsilon \to 0$.
2. *If we further assume that while the Riemannian manifolds M_i converge to Σ,* ***no collapse occurs*** *i.e. the Hausdorff dimension of Σ is the same as of M_i, then, the sectional curvature $K_i(p)$ of M_i at p satisfies,*
   $$\lim_{\epsilon \to 0} K_i(p) = \kappa;$$

The theorem above answers both questions 1 and 2. We can control the curvature of the obtained smooth signals in an adaptive way by making it converge to the $\limsup$ of Alexandrov curvature of the signal Σ.

3.2 Gluing

By arguments similar to those in [10] we have,

Theorem 10 ([1]) *Let the above smooth approximations of Σ be given in neighborhoods of two points p, q. Then they coincide as well as their sectional curvatures K_{i,V_i}, K_{i,U_i} on the common boundary, if non empty.*

4. Sampling of geometric signals

We propose the following scheme for sampling of a geometric signals.

1. Consider the signal as an Alexandrov space. This requires the representation of the signal as a tame metric space in a meaningful manner.
2. Assess the appropriate Alexandrov curvature bound. This can be done by the use of discrete metric curvature measures.
3. Smooth the signal while controlling the curvature of the smoothed signal to suitably approximate the estimated curvature.
4. Sample the smoothed signal according to Theorem 1

4.1 Special case - images

It is common to regard images as surfaces embedded in some $\mathbb{R}^n$. For gray scale images $\mathbb{R}^3$ is considered while for color images it is usual to take $\mathbb{R}^5$. Figure 2 shows image re-sampled according to the geometric sampling proposed in Theorem1. In this example no smoothing was applied prior to sampling and artifacts of this can be seen in the reconstructed image. "Flat areas" of the image have 20 times reduced sampling resolution with respect to the original resolution.

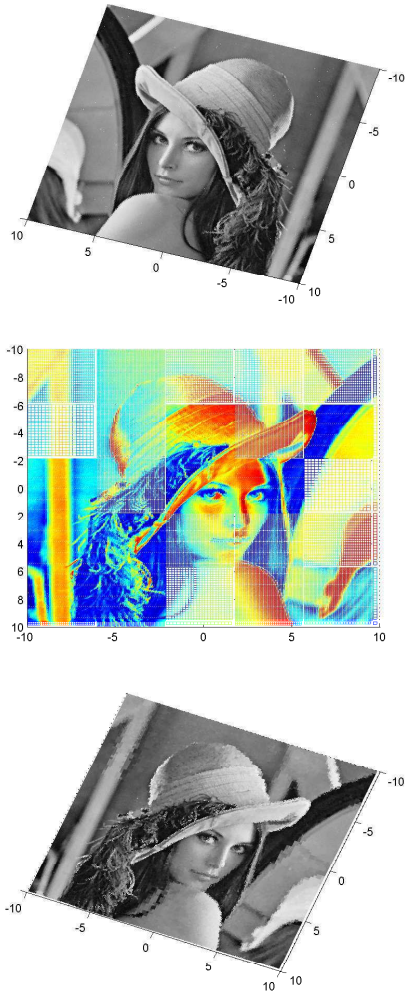


Figure 2: Geometric sampling of a gray scale image. **Top to bottom** - original Lena; Lena resampled. The white dots are the new sampling points. One can see the sparseness w.r.t the original; Lena reconstructed. Reconstruction using linear interpolation over the sampling points. No smoothing was done.

In order to estimate the curvature of an image as an Alexandrov space we can take the set of discrete curvature measures proposed in [5] where such measures are suggested for very general cell-complexes. It is shown in [5]

that the one-dimensional curvature measure resembles the **Ricci** curvature of a cell-complex which, in the case of images (since they are 2-dimensional manifolds) coincides with the Gaussian curvature. Figure 3 shows the combinatorial Ricci (= Gauss) curvature of the image in Figure 2, see [13] for details about the adoption of the curvature measures introduced in [5] to images.

Figure 3: Discrete Ricci curvature of Lena. Apart from giving an assessment for the curvature of the image as an Alexandrov space, it also serves as an excellent edge detector as itself.

5. Further study

Current and future studies of geometric sampling of images and signals, focus on two aspects. First we wish to modify the smoothing process introduced herein so it will be done in the Fourier domain rather than the spatial domain. Namely, we wish to smooth the Fourier transform of the signal while considering curvature in the Fourier plane. This is inspired by the Nash embedding Theorem [8] while the Fourier transform of a manifold is smoothen prior to its embedding thus achieving a higher degree of smoothness with respect to smoothing in the spatial domain.
Another direction of study is devoted to the development of a geometric theory of sparse representations and geometric compress sensing.

References:

[1] Appleboim, E., Saucan, E. and Zeevi, Y. Y. *Geometric reproducing kernels for signals*, preprint.

[2] Bochner, S. and Yano, K., *Curvature and Betti numbers*, Ann. Math. Stud. 32, 1953.

[3] Cheeger, J. and Gromov, M., *On the characteristic numbers of complete manifolds of bounded curvature and finite volume*, Diff. Geom. and Com. Anal. Chavel Farkas Ed., Springer, 1985.

[4] Cheeger, J. and Gromov, M., *Bounds on the Von Neumann dimension of $\mathbb{L}^2$-cohomology and the Gauss-Bonnet theorem for open manifolds*, J. Diff. Geom. 21, 1985.

[5] Forman, R., *Bochner's method for cell-complexes and combinatorial Ricci curvature*, Disc. Comp. Geom., 29, 2003.

[6] Kuwae, K. Machigashira, Y. and Shioya, T., *Sobolev spaces, Laplacian and heat kernel on Alexandrov spaces*, Math. Z. 238, 2001.

[7] Munkres, J. *Elementary Differential Topology*, Ann. Math. Stud. 54, 1966.

[8] Nash, J., *The Imbedding problem for Riemannian manifolds*, Ann. Math. 63, 1956.

[9] Otsu, Y. and Shioya, T., *The Riemannain stracture of Alexandrov Spaces*, J. Diff. Geom., 39, 1994.

[10] Petersen, P., Wei, G. and Ye, R., *Controlled geometry via smoothing*, Comm. Math. Helv., 74, 1999.

[11] Saucan, E., Appleboim, E. and Zeevi, Y. Y. *Sampling and Reconstruction of Surfaces and Higher Dimensional Manifolds*, J. Math. Imaging. Vis., 30, 2008.

[12] Saucan, E., Appleboim, E. and Zeevi, Y. Y. *Geometric Sampling of Manifolds for Image Representation and Processing* LNCS, 4485, 2007.

[13] Saucan, E. Appleboim, E., Wolansky G. and Zeevi, Y. Y., *Combinatorial Ricci curvature for image processing*, Midas Jour. Proc. MICCAI 2008

Multivariate Complex B-Splines, Dirichlet Averages and Difference Operators

Brigitte Forster [1,2] and Peter Massopust [2,1]

(1) Zentrum Mathematik, M6, Technische Universität München, Germany
(2) Institut für Biomathematik und Biometrie, Helmholtz Zentrum München, Germany
forster@ma.tum.de, massopust@ma.tum.de

Abstract:

For the Schoenberg B-splines, interesting relations between their functional representation, Dirichlet averages and difference operators are known. We use these relations to extend the B-splines to an arbitrary (infinite) sequence of knots and to higher dimensions. A new Fourier domain representation of the multidimensional complex B-spline is given.

1. Complex B-Splines

Complex B-splines are a natural extension of the classical Curry-Schoenberg B-splines [2] and the fractional splines first investigated in [16]. The complex B-splines $B_z : \mathbb{R} \to \mathbb{C}$ are defined in Fourier domain as

$$\mathcal{F}(B_z)(\omega) = \int_{\mathbb{R}} B_z(t) e^{-i\omega t}\, dt = \left(\frac{1-e^{-i\omega}}{i\omega}\right)^z$$

for $\operatorname{Re} z > 1$. They are well-defined, because of $\{\frac{1-e^{-i\omega}}{i\omega} \mid \omega \in \mathbb{R}\} \cap \{y \in \mathbb{R} \mid y < 0\} = \emptyset$ they live on the main branch of the complex logarithm. Complex B-splines are elements of $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. They have several interesting basic properties, which are discussed in [5]. Let $\operatorname{Re} z, \operatorname{Re} z_1, \operatorname{Re} z_2 > 1$.

- Complex B-splines B_z are piecewise polynomials of complex degree.
- Smoothness and decay:
  - $B_z \in W_2^r(\mathbb{R})$ for $r < \operatorname{Re} z - \frac{1}{2}$. Here $W_2^r(\mathbb{R})$ denotes the Sobolev space with respect to the L^2-Norm and with weight $(1+|x|^2)^r$.
  - $B_z(x) = \mathcal{O}(x^{-m})$ for $m < \operatorname{Re} z + 1$, $|x| \to \infty$.
- Recursion formula: $B_{z_1} * B_{z_2} = B_{z_1+z_2}$.
- Complex B-splines are scaling functions and generate multiresolution analyses and wavelets.
- But in general, they don't have compact support.
- Last but not least: They relate difference and differential operators.

In this paper, we take closer look at this last relation and the respective multivariate setting. To this end, we will consider the known relations between classical B-splines, difference operators and Dirichlet averages.

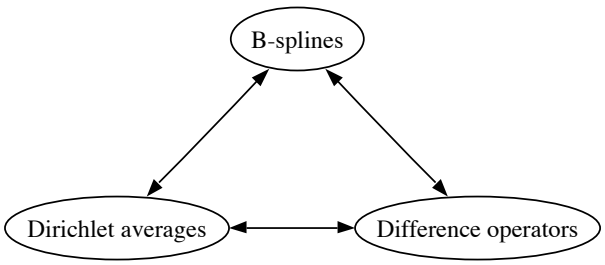


Figure 1: Relations between classical B-splines, difference operators and Dirichlet averages.

2. Representation in time-domain

We defined complex B-splines in Fourier domain, and Fourier inversion shows that these functions are piecewise polynomials of complex degree:

Proposition 1. [5] Complex B-splines have a time-domain representation of the form

$$B_z(t) = \frac{1}{\Gamma(z)} \sum_{k \geq 0} (-1)^k \binom{z}{k} (t-k)_+^{z-1},$$

pointwise for all $t \in \mathbb{R}$ and in $L^2(\mathbb{R})$-norm. Here,

$$t_+^z = \begin{cases} t^z = e^{z \ln t}, & \text{if } t > 0, \\ 0, & \text{if } t \leq 0, \end{cases}$$

is the truncated power function, and $\Gamma : \mathbb{C} \setminus \mathbb{Z}_0^- \to \mathbb{C}$ denotes the Euler Gamma function.

Compare: The cardinal B-spline B_n, $n \in \mathbb{N}$, has the similar representation

$$\begin{aligned} B_n(t) &= \frac{1}{(n-1)!} \sum_{k=0}^{n} (-1)^k \binom{n}{k} (t-k)_+^{n-1} \\ &= \frac{1}{\Gamma(n)} \sum_{k=0}^{\infty} (-1)^k \binom{n}{k} (t-k)_+^{n-1}. \end{aligned}$$

3. Relations to Difference Operators

It is well-known that in the construction of the Curry-Schoenberg B-splines difference operators are deeply involved. The same is true for complex B-splines. To establish the corresponding relation, let us first recall the definition of the backward difference operator ∇.

Let $g : \mathbb{R} \to \mathbb{C}$ be a function. Then the backward difference operator $\nabla = \nabla^1$ is recursively defined as follows:

$$\begin{aligned} \nabla g(t) &= g(t) - g(t-1), \\ \nabla^{n+1} g(t) &= \nabla(\nabla^n g(t)) \quad \text{for } n \in \mathbb{N}. \end{aligned}$$

This definition yields the explicit representation

$$\nabla^n g(t) = \sum_{k=0}^{n} \binom{n}{k} (-1)^k g(t-k).$$

For the cardinal B-splines B_n we can write:

$$\begin{aligned} B_n(t) &= \frac{1}{(n-1)!} \sum_{k=0}^{n} (-1)^k \binom{n}{k} (t-k)_+^{n-1} \\ &= \frac{1}{(n-1)!} \nabla^n t_+^{n-1}. \end{aligned}$$

In comparison: For the complex B-splines, we have an analog representation:

$$B_z(t) = \frac{1}{\Gamma(z)} \sum_{k=0}^{\infty} (-1)^k \binom{z}{k} (t-k)_+^{z-1}, \quad \operatorname{Re} z \geq 1.$$

This invites to define a complex difference operator:

Definition 2. [5, 6] The difference operator ∇^z of complex order z is defined as

$$\nabla^z g(t) := \sum_{k=0}^{\infty} (-1)^k \binom{z}{k} g(t-k), \quad z \in \mathbb{C}, \quad \operatorname{Re} z \geq 1.$$

Hence a second time domain representation of the complex B-spline is

$$B_z(t) = \frac{1}{\Gamma(z)} \nabla^z t_+^{z-1}.$$

In a similar way, we can establish a relation to divided differences. Recall that for a knot sequences $\{t_0, \ldots, t_n\} \subset \mathbb{R}$, $n \geq 1$, divided differences are recursively defined as follows. Let $g : \mathbb{R} \to \mathbb{C}$ be some function.

$$\begin{aligned} [t_0]g &= g(t_0), \\ [t_0, \ldots, t_n]g &= \frac{[t_0, \ldots, t_{n-1}]g - [t_1, \ldots, t_n]g}{t_0 - t_n} \\ &= \sum_{j=0}^{n} \frac{g(t_j)}{\prod_{l \neq j} (t_j - t_l)}. \end{aligned}$$

For the cardinal B-spline,

$$\begin{aligned} B_n(t) &= \frac{1}{(n-1)!} \sum_{k=0}^{n} (-1)^k \binom{n}{k} (t-k)_+^{n-1} \\ &= n \sum_{k=0}^{n} (-1)^k \frac{1}{k!(n-k)!} (t-k)_+^{n-1} \\ &= (-1)^n n \sum_{k=0}^{n} \frac{(t-k)_+^{n-1}}{\prod_{l \neq k} (k-l)} \\ &= (-1)^n n [0, 1, \ldots, n](t - \bullet)_+^{n-1}. \end{aligned}$$

(The factor $(-1)^n$ is due to our representation of the cardinal B-spline via backward difference operators.)
The same ideas give rise to the definition of complex divided differences.

Definition 3. Let $g : \mathbb{R} \to \mathbb{C}$ be some function. We define the complex divided differences for the knot sequence $\mathbb{N}_0$ via

$$[z; \mathbb{N}_0]g := \sum_{k \geq 0} (-1)^k \frac{g(k)}{\Gamma(z-k+1)\Gamma(k+1)}.$$

Then the complex B-spline can be written as

$$B_z(t) = z[z, \mathbb{N}_0](t - \bullet)_+^{z-1}.$$

Comparing "old" and "new" divided difference operator for $z = n \in \mathbb{N}$, yields

$$(-1)^n [0, 1, \ldots n] = [n, \mathbb{N}_0].$$

Proposition 4. [6, 7] Let $\operatorname{Re} z > 0$ and $g \in \mathcal{S}(\mathbb{R}^+)$. Then

$$[z; \mathbb{N}_0]g = \frac{1}{\Gamma(z+1)} \int_{\mathbb{R}} B_z(t) g^{(z)}(t)\, dt,$$

where $g^{(z)} = W^z g$ is the complex Weyl derivative:
For $n = \lceil \operatorname{Re} z \rceil$, $\nu = n - z$,

$$W^z g(t) = (-1)^n \frac{d^n}{dt^n} \left[\frac{1}{\Gamma(\nu)} \int_t^{\infty} (x-t)^{\nu-1} g(x)\, dx \right].$$

Sketch of proof:

$$\begin{aligned} & \frac{1}{\Gamma(z+1)} \int_{\mathbb{R}} B_z(t) g^{(z)}(t)\, dt \\ = \; & \frac{1}{\Gamma(z+1)} \int_{\mathbb{R}} z[z, \mathbb{N}_0](t - \bullet)_+^{z-1} W^z g(t)\, dt \\ = \; & [z, \mathbb{N}_0] \frac{1}{\Gamma(z)} \int_{\bullet}^{\infty} (t - \bullet)_+^{z-1} W^z g(t)\, dt \\ = \; & [z, \mathbb{N}_0] W^{-z} W^z g \quad = \quad [z, \mathbb{N}_0] g. \end{aligned}$$

Here, $W^{-z} f = \frac{1}{\Gamma(z)} \int_{\bullet}^{\infty} (t - \bullet)_+^{z-1} f(t)\, dt$ is the complex Weyl integral of the function f, i.e., the inverse operator of W^z. □
Now we are able to establish a first relation between divided difference operators and Dirichlet averages.

Proposition 5. (Generalized Hermite-Genocchi-Formula: Divided Differences and Dirichlet Averages) [6, 7]
Let Δ^∞ be the infinite-dimensional simplex

$$\Delta^\infty := \{u := (u_j) \in (\mathbb{R}_0^+)^{\mathbb{N}_0} | \sum_{j=0}^{\infty} u_j = 1\} = \varprojlim \Delta^n,$$

defined as the projective limit of the finite dimensional simplices Δ^n, and let μ_e^∞ be the generalized Dirichlet measure defined by the projective limit

$$\mu_e^\infty = \varprojlim \Gamma(n+1) \lambda^n,$$

where λ^n the Lebesgue measure on Δ^n. Then

$$\begin{aligned} [z, \mathbb{N}_0]g &= \frac{1}{\Gamma(z+1)} \int_{\Delta^\infty} g^{(z)}(\mathbb{N}_0 \cdot u) d\mu_e^\infty(u) \\ &= \frac{1}{\Gamma(z+1)} \int_{\mathbb{R}} B_z(t) g^{(z)}(t)\, dt \end{aligned}$$

for all real-analytic $g \in \mathcal{S}(\mathbb{R}^+)$.

Up to now we have considered complex B-splines with knot sequence $\mathbb{N}_0$ and derived from there new difference operators and finally the relation to Dirichlet averages, just as indicated in the diagram in Fig. 1:

B-splines $\rightarrow$ Difference operators $\rightarrow$ Dirichlet averages.

Our next step will consist of generalizing the setting with appropriate weights in travelling through the diagram another way round: Dirichlet averages for other knot sequences τ and with weights $\rightarrow$ Generalized B-splines with knot sequence $\tau \rightarrow$ Difference operators.

4. Splines and Dirichlet Averages

Let $b \in \mathbb{R}_+^\infty$ be a weight vector and $\tau = \{t_k\}_{k\in\mathbb{N}_0} \in \mathbb{R}_+^{\mathbb{N}_0}$ an increasing sequence of knots with $\limsup_{k\to\infty} \sqrt[k]{t_k} \leq \rho < e$.

Definition 6. A complex B-spline $B_z(\bullet \mid b;\tau)$ with weight vector b and knot sequence τ is a function satisfying

$$\int_{\mathbb{R}} B_z(t \mid b;\tau) g^{(z)}(t)\, dt = \int_{\Delta^\infty} g^{(z)}(\tau\cdot u) d\mu_b^\infty(u) \quad (1)$$

for all real-analytic $g \in \mathcal{S}(\mathbb{R}^+)$. Here, $\mu_b^\infty = \varprojlim \mu_b^n$ is the projective limit of Dirichlet measures with densities

$$\frac{\Gamma(b_0)\dots\Gamma(b_n)}{\Gamma(b_0+\ldots+b_n)} u_0^{b_0-1} u_1^{b_1-1}\dots u_n^{b_n-1}.$$

Since both W^z and W^{-z} are linear operators mapping $\mathcal{S}(\mathbb{R}^+)$ into itself [11, 15] and since the real-analytic functions in $\mathcal{S}(\mathbb{R}^+)$ are dense in $\mathcal{S}(\mathbb{R}^+)$ [13], (1) holds for all $g \in \mathcal{S}(\mathbb{R}^+)$. Moreover, since $\mathcal{S}(\mathbb{R}^+)$ is dense in $L^2(\mathbb{R}^+)$, we deduce that $B_z(\bullet \mid b,\tau) \in L^2(\mathbb{R}^+)$.
Equation (1) means, we define the weighted version of the complex B-spline in a weak sense via Dirichlet averages. Referring again to the diagram in Fig. 1, we now move from the generalized B-splines to generalized divided differences.

Definition 7. For knot sequences $\tau \in \mathbb{R}_+^{\mathbb{N}_0}$ and weight vectors $b \in \mathbb{R}_+^\infty$ as above, we define the generalized complex divided differences $[z;\tau]_b$ as follows. Let $g : \mathbb{R} \to \mathbb{C}$ be some function.

$$[z;\tau]_b g := \frac{1}{\Gamma(z)} \int_{\mathbb{R}} B_z(t|b;\tau) g^{(z)}(t)\, dt$$

for all $g \in \mathcal{S}(\mathbb{R})$.

This definition is compatible with the usual Dirichlet splines. In fact, for all finite $\tau = \tau(n) \in \mathbb{R}_+^{n+1}$ and $b = b(n) \in \mathbb{R}_+^{n+1}$, and for $z = n \in \mathbb{N}_0$ the Dirichlet spline $D_n(\bullet|b;\tau)$ of order n is defined by

$$\begin{aligned}\int_{\mathbb{R}} g^{(n)}(t) D_n(t|b;\tau)\, dt &= \int_{\Delta^n} g^{(n)}(\tau\cdot u)\, d\mu_b^n(u)\\ &= G^{(n)}(b;\tau)\end{aligned}$$

for all $g \in C^n(\mathbb{R})$. Here, G is the Dirichlet average of g:

$$G(b;\tau) = \int_{\Delta^n} g(\tau\cdot u)\, d\mu_b^n(u).$$

5. Multivariate Complex B-Splines

To define complex B-splines in a multivariate setting, we consider ridge functions and define multivariate B-splines on their basis. Then, we walk again through the diagram in Fig. 1: Multivariate B-splines $\rightarrow$ Multivariate difference operators. Results on Dirichlet averages yield new recurrence relations for multivariate B-splines: Dirichlet averages $\rightarrow$ B-splines.
Note that the approach via ridge functions had already let to an extension of the Curry-Schoenberg-splines to a multivariate setting, e.g. [3, 4, 10, 12]. However, some of these approaches have certain restrictions on the knots and none of them considers complex splines.
Given $\lambda \in \mathbb{R}^s\backslash\{0\}$, a direction, and $g : \mathbb{R} \to \mathbb{C}$ a function. The ridge function g_λ corresponding to g is defined via

$$g_\lambda : \mathbb{R}^s \to \mathbb{C}, \quad g_\lambda(x) = g(\langle\lambda, x\rangle) \quad \text{for all } x \in \mathbb{R}^s.$$

Definition 8. [9] Let $\boldsymbol{\tau} = \{\tau^n\}_{n\in\mathbb{N}_0} \in (\mathbb{R}^s)^{\mathbb{N}_0}$ a sequence of knots in $\mathbb{R}^s$ with $\limsup_{n\to\infty} \sqrt[n]{\|\tau^n\|} \leq \rho < e$. The multivariate complex B-spline $\boldsymbol{B}_z(\bullet|b;\boldsymbol{\tau})$ with weights $b \in \mathbb{C}_+^{\mathbb{N}_0}$ and knots $\boldsymbol{\tau}$ is defined on ridge functions via

$$\int_{\mathbb{R}^s} g(\langle\lambda,x\rangle)\boldsymbol{B}_z(x \mid b;\boldsymbol{\tau})\, dx = \int_{\mathbb{R}} g(t) B_z(t \mid b;\lambda\boldsymbol{\tau})\, dt, \quad (2)$$

where $g \in \mathcal{S}(\mathbb{R}^+)$ and $\lambda \in \mathbb{R}^s \setminus \{0\}$, such that $\lambda\boldsymbol{\tau} = \{\langle\lambda,\tau^n\rangle\}_{n\in\mathbb{N}_0}$ is separated.

Since ridge functions are dense in $L^2(\mathbb{R}^s)$ [14], we deduce that $\boldsymbol{B}_z(\bullet \mid b;\boldsymbol{\tau}) \in L^2((\mathbb{R}^+)^s)$.

Example 9. (Divided differences in the multivariate case) Given $b = e := (1,1,1,\ldots)$. Then for all $g \in \mathcal{S}(\mathbb{R}^\infty)$:

$$\begin{aligned}[z;\boldsymbol{\tau}]_e g_\lambda &= [z;\boldsymbol{\tau}]g_\lambda \;=\; [z;\boldsymbol{\tau}]g(\langle\lambda,\bullet\rangle)\\ &= \frac{1}{\Gamma(z)}\int_{\mathbb{R}^s} g^{(z)}(\langle\lambda,x\rangle)\boldsymbol{B}_z(x \mid e;\boldsymbol{\tau})\, dx\\ &= \frac{1}{\Gamma(z)}\int_{\mathbb{R}} g^{(z)}(t) B_z(t \mid e;\lambda\boldsymbol{\tau})\, dt = [z;\lambda\boldsymbol{\tau}]g.\end{aligned}$$

for all $\lambda \in \mathbb{R}^s$ such that $\lambda\boldsymbol{\tau}$ is separated.

Example 10. (Multivariate cardinal B-splines) For $n \in \mathbb{N}$ and a finite sequence of knots $\boldsymbol{\tau} = \{\tau^0,\tau^1,\ldots,\tau^n\}$:

$$\begin{aligned}&[\tau^0,\ldots,\tau^n]g_\lambda := [n;\boldsymbol{\tau}]g(\langle\lambda,\bullet\rangle)\\ =\;& \frac{1}{n!}\int_{\mathbb{R}^s} g^{(n)}(\langle\lambda,x\rangle)\boldsymbol{B}_n(x \mid e;\boldsymbol{\tau})\, dx\\ =\;& \frac{1}{n!}\int_{\mathbb{R}} g^{(n)}(t) B_n(t \mid e;\lambda\boldsymbol{\tau})\, dt\\ =\;& [n;\lambda\boldsymbol{\tau}]g = \sum_{j=0}^n \frac{g(\langle\lambda,\tau^j\rangle)}{\prod_{l\neq j}\langle\lambda,\tau^j-\tau^l\rangle}.\end{aligned}$$

Given a sequence of knots $\boldsymbol{\tau} \subset \mathbb{R}^s$ and a weight vector b as above. In addition, let $b \in l^1(\mathbb{N}_0)$ such that $\|b\|_1 =: c$. Then the Dirichlet averages of $g^{(z)} \in \mathcal{D}(\mathbb{R})$ and $g_j^{(z+1)} := (\langle\lambda,\tau^j\rangle - \bullet)g^{(z+1)}$, $j \in \mathbb{N}_0$, satisfy:

$$(c-1)G^{(z)}(b;\lambda\boldsymbol{\tau}) = (c-1)G^{(z)}(b-e_j;\lambda\boldsymbol{\tau}) + G_j^{(1+z)}(b;\lambda\boldsymbol{\tau}).$$

For the finite dimensional case see [1, 12]. These and other relations of similar type on Dirichlet averages yield new results for multivariate complex B-splines. As a example, we state:

Proposition 11. [9] Under the above conditions, for all $j \in \mathbb{N}_0$:

$$\begin{aligned}
&(c-1)\int_{\mathbb{R}^s} g_\lambda^{(z)}(x)\boldsymbol{B}_z(x \mid b;\boldsymbol{\tau})dx = \\
= \quad &(c-1)\int_{\mathbb{R}^s} g_\lambda^{(z)}(x)\boldsymbol{B}_z(x \mid b-e_j;\boldsymbol{\tau})dx \\
+ \quad &\int_{\mathbb{R}^s} \langle \lambda, \tau^j - x\rangle\, g_\lambda^{(1+z)}(x)\boldsymbol{B}_z(x \mid b;\boldsymbol{\tau})dx.
\end{aligned}$$

More relations of this type are given in [8].

6. Fourier representation of multivariate complex B-splines

We saw above that both the univariate and the multivariate complex B-splines are L^2-functions: $B_z(\bullet \mid b;\tau) \in L^2(\mathbb{R}^+)$ and $\boldsymbol{B}_z(\bullet \mid b;\boldsymbol{\tau}) \in L^2((\mathbb{R}^+)^s)$ Therefore, we can apply the Plancherel transform to both functions and consider their frequency spectrum.
Let $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_s) \in \mathbb{R}^s$ and let $\lambda \in \mathbb{R}^s$, $\|\lambda\| = 1$, be the direction of $\boldsymbol{\omega}$, i.e., $\boldsymbol{\omega} = \omega\lambda$ for some $\omega \geq 0$. For the Fourier transform of the generalized complex B-spline we have for $x = (x_1, \ldots, x_s) \in \mathbb{R}^s$:

$$\begin{aligned}
&\widehat{B}_z(\omega \mid b;\lambda\boldsymbol{\tau}) = \\
&= \int_{\mathbb{R}} e^{-i\omega t} B_z(t \mid b;\lambda\boldsymbol{\tau})\, dt \\
&= \int_{\mathbb{R}^s} e^{-i\omega\langle\lambda,x\rangle} \boldsymbol{B}_z(x \mid b;\boldsymbol{\tau})\, dx \\
&= \int_{\mathbb{R}^s} e^{-i\omega(\lambda_1 x_1 + \ldots + \lambda_s x_s)} \boldsymbol{B}_z(x \mid b;\boldsymbol{\tau})\, dx \\
&= \int_{\mathbb{R}^s} e^{-i(\omega_1 x_1 + \ldots + \omega_s x_s)} \boldsymbol{B}_z(x \mid b;\boldsymbol{\tau})\, dx \\
&= \int_{\mathbb{R}^s} e^{-i\langle\boldsymbol{\omega},\boldsymbol{x}\rangle} \boldsymbol{B}_z(x \mid b;\boldsymbol{\tau})\, dx \\
&= \widehat{\boldsymbol{B}}_z(\boldsymbol{\omega} \mid b;\boldsymbol{\tau}) \quad = \quad \widehat{\boldsymbol{B}}_z(\omega\lambda \mid b;\boldsymbol{\tau}).
\end{aligned}$$

This shows that the frequency spectrum of the multivariate complex B-spline along directions λ is given by the spectrum of the univariate spline with knots projected onto these λ.

7. Summary

Complex B-splines allow to define difference and divided difference operators of complex order for arbitrary knots and weights. Via their relation to Dirichlet averages and Dirichlet splines, they can be extended to higher dimensions via ridge functions. The Fourier transform of the univariate and multivariate complex B-spline are also related on ridges.

8. Acknowledgments

This work was partially supported by the grant MEXT-CT-2004-013477, Acronym MAMEBIA, of the European Commission.

References:

[1] B. C. Carlson. B-Splines, hypergeometric functions and Dirichlet averages. *J. Approx. Th.*, 67:311–325, 1991.

[2] H. B. Curry and I. J. Schoenberg. On spline distributions and their limits: The Pólya distribution functions. *Bulletin of the AMS*, 53(7–12):1114, 1947. Abstract.

[3] W. Dahmen and C. A. Micchelli. Statistical Encounters with B-Splines. *Contemporary Mathematics*, 59:17–48, 1986.

[4] C. de Boor. Splines as linear combinations of B-splines. In G. G. Lorentz et al., editor, *Approximation Theory II*, pages 1–47. Academic Press, 1976.

[5] B. Forster, T. Blu, and M. Unser. Complex B-splines. *Appl. Comp. Harmon. Anal.*, 20:281–282, 2006.

[6] B. Forster and P. Massopust. Statistical encounters with complex B-Splines. to appear in Constructive Approximation.

[7] B. Forster and P. Massopust. Some remarks about the connection between fractional divided differences, fractional B-Splines, and the Hermite-Genocchi formula. *International Journal of Wavelets, Multiresolution and Information Processing*, 6(2):279–290, 2008.

[8] P. Massopust. Double Dirichlet averages and complex B-splines. Submitted to SAMPTA 2009.

[9] P. Massopust and B. Forster. Multivariate complex B-splines and Dirichlet averages. Submitted to Journal of Approximation Theory.

[10] C. A. Micchelli. A constructive approach to Kergin interpolation in $\mathbb{R}^k$: Multivariate B-splines and Lagrange interpolation. *Rocky Mt. J. Math.*, 10(3):485–497, 1980.

[11] K. S. Miller and B. Ross. *An introduction to the fractional calculus and fractional differential equations*. Wiley, 1993.

[12] E. Neuman and P. J. Van Fleet. Moments of Dirichlet splines and their applications to hypergeometric functions. *Journal of Computational and Applied Mathematics*, 53:225–241, 1994.

[13] O. V. Odinokov. Spectral analysis in certain spaces of entire functions of exponential type and its applications. *Izv. Math.*, 64(4):777–786, 2000.

[14] A. Pinkus. Approximating by ridge functions. In A. Le Méhauté, C. Rabut, and L. L. Schumaker, editors, *Surface Fitting and Multiresolution Methods*, pages 1–14. Vanderbilt University Press, 1997.

[15] S. G. Samko, A. A. Kilbas, and O. I. Marichev. *Fractional Integrals and Derivatives*. Gordon and Breach Science Publishers, Minsk, Belarus, 1987.

[16] M. Unser and T. Blu. Fractional splines and wavelets. *SIAM Review*, 42(1):43–67, March 2000.

Concrete and discrete operator reproducing formulae for abstract Paley–Wiener space

J.R. Higgins

I.H.P., 4 rue du Bary, 11250 Montclar, France.
rowlandhiggins@yahoo.com

Abstract:

The classical Paley–Wiener space possesses two reproducing formulae; a 'concrete' reproducing equation and a 'discrete' analogue, or sampling series, and there is a striking comparison between them. It is shown that such analogies persist in the setting of Paley–Wiener spaces that are more general than the classical case. In fact, there are 'operator' versions of the reproducing equation and of the sampling series that are also comparable, not 'exactly' but nearly so. Reproducing kernel theory and abstract harmonic analysis are brought together to achieve this, then the special case of multiplier operators with respect to the Fourier transform is considered. The Riesz transforms provide a two-dimensional example, with possibilities of extension to higher dimensions and to further classes of operators.

1. Introduction

It has often been remarked that the classical Paley–Wiener space possesses two reproducing formulae; a 'concrete' reproducing equation

$$f(s) = \int_{\mathbb{R}} f(t)\,\mathrm{sinc}(s-t)dt, \quad (s \in \mathbb{R}), \qquad (1)$$

and a 'discrete' reproducing equation, or sampling series,

$$f(s) = \sum_{n\in\mathbb{Z}} f(n)\ \mathrm{sinc}(s-n), \quad (s \in \mathbb{R}), \qquad (2)$$

and that there is a striking analogy between the two (see, e.g., [3, p. 58]). Here, sinc denotes the standard function $\mathrm{sinc}\,x := (\sin \pi x)/\pi x$.

The purpose of the present lecture is to point out that concrete and discrete reproducing formulae and analogies between them persist in the setting of Paley–Wiener spaces that are more general than the classical case. It will be shown that for suitably chosen operators there are 'operator' versions of the reproducing equation and of the sampling series that are also comparable, in the same way as in the classical case described above.

2. The setting

Abstract theories that lead to reproducing formulae are outlined in §2.1 and §2.2, and are brought together in §2.3.

2.1 The reproducing kernel theory

The basic setting of this paper is that of the reproducing kernel theory of Saitoh [8, Ch. 2, §1]. Very briefly the background is as follows. Let E be an abstract set. For each t belonging to E let K_t belong to H (a separable Hilbert space with inner product denoted by $\langle , \rangle_H$). Then $k(s,t) := \langle K_t, K_s\rangle_H$ is defined on $E \times E$ and is called the *kernel function* of the map K_t. This kernel function is a *positive matrix* [8, Ch. 2, §2] and as such it determines one and only one Hilbert space for which it is the reproducing kernel. This Hilbert space is denoted by $R(\mathcal{K})$ since it turns out to be the set of images of H under the transformation $(\mathcal{K}g)(t) := \langle g, K_t\rangle_H$, $(g \in H)$.

Theorem 1 (Saitoh) *With the notations established above, $R(\mathcal{K})$ (which is now abbreviated to just R) is a Hilbert space which has the reproducing kernel $k(\cdot,\cdot)$, and is uniquely determined by this kernel k. For $f \in R$ there exists $\alpha \in H$ such that*

$$\|f\|_R = \|\mathcal{K}\alpha\|_R \le \|\alpha\|_H, \qquad (3)$$

and there exists a unique member, g say, of the class of all α's satisfying (3) such that

$$f(t) = \langle g, K_t\rangle_H, \qquad (t \in E),$$

and

$$\|f\|_R = \|g\|_H.$$

The reproducing equation for $f \in R$ is

$$f(t) = \langle f, k(\cdot,t)\rangle_R \qquad (4)$$

The following theorem is simple but very useful.

Theorem 2 *The convergence of a sequence in the norm of R implies that it converges pointwise over E, and the convergence is uniform over any subset of E on which $k(t,t) = \|k(\cdot,t)\|^2$ is bounded.*

The following Theorem is to be found in [8].

Theorem 3 *With notations as above, let $\{s_n\}$, $(n \in \mathbb{X})$, be points of E such that $\{K_{s_n}\}$ is an orthonormal basis for H. Then the sampling series representation*

$$f(t) = \sum_{n\in\mathbb{X}} f(s_n)k(s_n,t), \qquad (5)$$

holds, convergence being in the norm of R; and then of course Theorem 2 applies.

2.2 Abstract harmonic analysis

A very brief introduction (mostly just notations) to the abstract harmonic analysis that will be needed is now given. All necessary background, and much more, is to be found in [1], [2].

Let G be a locally compact abelian (LCA) group (written additively). Let (t, γ) be a character of G, that is, a continuous homomorphism of G into the circle group. Let $G^\wedge = \Gamma$ denote the group of continuous characters on G, usually called the dual group of G. We assume that Γ has a countable discrete subgroup Λ.

Haar measures on the various groups are normalised in the standard way [1], and this means in particular that there is a measurable transversal (i.e., a complete set of coset representatives) $\Omega \subset \Gamma$ of Γ/Λ, and it has finite Haar measure.

Now

$$\mathfrak{H} = \Lambda^\perp := \{t \in G : (t, \lambda) = 1, (\lambda \in \Lambda)\}.$$

is a subgroup of G and is called the 'annihilator' of Λ. We assume that $\mathfrak{H}$ is discrete; it follows that the quotient group Γ/Λ is compact.

The Fourier transform on $L^2(G)$ is defined in the usual way:

$$f^\wedge(\gamma) = (\mathcal{F}f)(\gamma) := \int_G f(t)(t, \gamma)\, dt,$$

in the L^2 sense, where dt denotes the element of Haar measure on G (likewise, $d\gamma$ denotes the element of Haar measure on Γ). The inverse Fourier transform will be denoted by $^\vee$ or by $\mathcal{F}^{-1}$.

We shall need the 'shift' property of the Fourier transform: $f(\cdot - x)^\wedge(\gamma) = (-x, \gamma) f^\wedge(\gamma)$.

Abstract Paley Wiener space $PW_\Omega(G)$ is defined as follows:

$$PW_\Omega(G) := \{f : f \in L^2(G) \cap C(G), \quad f^\wedge(\gamma) = 0 \quad (Haar) \ \ a.a.\ \gamma \notin \Omega\} \quad (6)$$

2.3 Combining harmonic analysis with the reproducing kernel theory

The abstract set E of §2.1 is often taken to be $\mathbb{R}$ or $\mathbb{C}$. Here, however, we take it to be an LCA group G thus combining two abstract theories, harmonic analysis and the reproducing kernel theory. In the notations of §2.1 and §2.2 we also take $K_t = (t, \cdot)$, $H = L^2(\Omega)$ and $\mathcal{K}g = \mathcal{F}^{-1}g$, $g \in L^2(\Omega)$. Then we have

$$\begin{aligned} k(s,t) &= \int_\Omega (t, \gamma)\overline{(s, \gamma)}\, d\gamma = \int_\Omega (t - s, \gamma)\, d\gamma \\ &= \big(\chi_\Omega{}^\vee\big)(t - s) =: \varphi_\Omega(t - s), \end{aligned} \quad (7)$$

where χ_S denotes the characteristic function of a set S. It does not seem to have been recognised that $\varphi_\Omega(t - s)$ is the reproducing kernel for $PW_\Omega(G)$, and that this allows a close association between sampling in the harmonic analysis setting and Saitoh's theory.

The space R of §2.1 is now seen to be the Paley–Wiener space defined in (6), and its reproducing equation is

$$f(t) = \langle f, \varphi_\Omega(t - \cdot)\rangle_{L^2(G)} \quad (8)$$

Kluvánek's sampling theorem [4, p. 45] is a consequence:

Theorem 4 *Let $f \in PW_\Omega$. With the assumptions of §2.2,*

$$f(t) = \sum_{h \in \mathfrak{H}} f(h)\varphi_\Omega(t - h) \quad (9)$$

in norm, etc., (see Theorem 2).

Our concrete – discrete comparison is beween (8) and (9).

3. Operator kernels and operator reproducing formulae

The presence of kernels and reproducing equations associated with operators on a reproducing kernel Hilbert space add greatly to the richness of its structure, as we shall see in this section.

3.1 Operator kernels and operator reproducing equations

Let R be the separable Hilbert space of functions defined on E with reproducing kernel $k(s, t)$, as we have discussed it in §2.1. Let $\mathcal{B}$ be a bijection on R, and let $\mathcal{B}^*$ denote the adjoint operator. The action of $\mathcal{B}$ on R is governed by the action of $\mathcal{B}^*$ on the reproducing kernel k, because for $f \in R$,

$$\big(\mathcal{B}f\big)(t) = \langle \mathcal{B}f, k(\cdot, t)\rangle_R = \langle f, \mathcal{B}^* k(\cdot, t)\rangle_R. \quad (10)$$

See, e.g., [5].

Definition 1 *The kernel*

$$h(s,t) := \big(\mathcal{B}^* k(\cdot, t)\big)(s), \qquad s, t \in E$$

will be called the operator kernel *of $\mathcal{B}$.*

In this notation (10) is

$$\big(\mathcal{B}f\big)(t) = \langle f, h(\cdot, t)\rangle. \quad (11)$$

Now from Definition 1 above, $((\mathcal{B}^*)^{-1}h(\cdot, t))(s) = k(s, t)$, so that, using the ordinary reproducing formula (4), we have

$$\begin{aligned} f(t) &= \langle f, k(\cdot, t)\rangle = \langle f, (\mathcal{B}^*)^{-1}h(\cdot, t)\rangle \\ &= \langle \big((\mathcal{B}^*)^{-1}\big)^* f, h(\cdot, t)\rangle. \end{aligned}$$

Now using standard properties of operators and their adjoints (e.g., [6, p. 202]) we can summarise these calculations as:

$$f(t) = \langle (\mathcal{B}^{-1}f)(\cdot), h(\cdot, t)\rangle. \quad (12)$$

This formula tells us that f can be reproduced, not from its own values as in the ordinary reproducing kernel theory, but from the result of acting on it with an operator. We can call this an *operator reproducing equation* in analogy with the ordinary reproducing equation (4).

Similar formulae for $\mathcal{B}^*$ can be obtained in the same way. First, we make the following

Definition 2

$$h^*(s,t) := \overline{h(t,s)} \qquad ((t,s) \in E \times E)$$

will be called the adjoint operator kernel *of* $\mathcal{B}$.

Kernels and their adjoints occur in important areas of study such as the theory of integral equations (see, e.g., [6, p. 170] for basic information). We shall find series expansions for such kernels and identify the action of h^* explicitly in Theorem 5 below.
First, let $\{\varphi_n\}$, $n \in \mathbb{X}$, be an orthonormal basis for R.

Lemma 1

$$h(s,t) = \sum_{n\in\mathbb{X}} \overline{\big(\mathcal{B}\varphi_n\big)(t)}\, \varphi_n(s), \qquad (s,t \in E). \quad (13)$$

Convergence is in the norm of R *for each* $t \in E$*, and the pointwise convergence is governed by Theorem 2.*

Proof The coefficients for the expansion of $h(\cdot,t)$, t fixed, in the basis $\{\varphi_n\}$ are

$$\langle h(\cdot,t), \varphi_n\rangle = \overline{\langle \varphi_n, h(\cdot,t)\rangle} = \overline{\big(\mathcal{B}\varphi_n\big)(t)}$$

by (11), thus (13) is obtained. □
It will be recalled that if we put

$$\begin{cases} \mathcal{B}\varphi_n = \psi_n \\ (\mathcal{B}^*)^{-1}\varphi_n = \psi_n^*, \end{cases} \quad (14)$$

then $\{\psi_n\}$ is a Riesz basis for R with dual basis $\{\psi_n^*\}$. In this notation (13) can be written

$$h(s,t) = \sum_{n\in\mathbb{X}} \overline{\psi_n(t)}\varphi_n(s). \quad (15)$$

Hence by Definition 2 we have

$$h^*(s,t) = \sum_{n\in\mathbb{X}} \overline{\varphi_n(t)}\psi_n(s). \quad (16)$$

in the norm of R for each $t \in E$.
By uniqueness the coefficients $\{\overline{\varphi_n(t)}\}$ are such that

$$\varphi_n(t) = \overline{\langle h^*(\cdot,t), \psi_n^*\rangle} = \langle (\mathcal{B}^*)^{-1}\varphi_n, h^*(\cdot,t)\rangle, \quad (17)$$

Since this relationship is true for every member φ_n of a basis for R, it holds for every $f \in R$ by the usual density argument. This argument runs as follows:
Let $\sum_N c_n\varphi_n$ be the Nth partial sum of the expansion for f in the basis φ_n. Then taking linear combinations in (17),

$$\sum_N c_n\varphi_n(t) = \langle (\mathcal{B}^*)^{-1} \sum_N c_n\varphi_n(t), h^*(\cdot,t)\rangle. \quad (18)$$

Consider

$$f(t) - \langle (\mathcal{B}^*)^{-1} f, h^*(\cdot,t)\rangle. \quad (19)$$

Put $f(t) - \sum_N c_n\varphi_n(t) = F_n(t)$. Now inserting the right and left hand sides of (18) we find from (19) that

$$\begin{aligned} &\big|f(t) - \langle (\mathcal{B}^*)^{-1} f, h^*(\cdot,t)\rangle\big| \quad (20)\\ &= \big|F_n(t) - \langle (\mathcal{B}^*)^{-1}(f - \sum_N c_n\varphi_n), h^*(\cdot,t)\rangle\big| \\ &\le \big|F_n(t)\big| + \big|\langle (\mathcal{B}^*)^{-1}(f - \sum_N c_n\varphi_n), h^*(\cdot,t)\rangle\big| \\ &\le \big|F_n(t)\big| + \|(\mathcal{B}^*)^{-1}(f - \sum_N c_n\varphi_n)\|\|h^*(\cdot,t)\| \\ &\le \big|F_n(t)\big| + B\|f - \sum_N c_n\varphi_n\|\|h^*(\cdot,t)\| \quad (21) \end{aligned}$$

for a constant B which is consequent upon the fact that, since $\mathcal{B}$ is bounded, $\mathcal{B}^*$ is bounded and by Banach's 'bounded inverse' theorem $(\mathcal{B}^*)^{-1}$ is bounded.
Now N can be made to approach ∞. Since $F_n(t)$ converges to 0 both in norm and pointwise on E (see Theorem 2), the expression in (21) approaches 0 for each fixed $t \in E$. Finally, from (20) we obtain the following

Theorem 5 *Let* R*,* $\mathcal{B}$ *and* E *be as above. Then we have the adjoint operator reproducing formula*

$$f(t) = \langle (\mathcal{B}^*)^{-1} f, h^*(\cdot,t)\rangle.$$

This shows the basic property of h^*; it reproduces f from $(\mathcal{B}^*)^{-1}f$.

3.2 Operator sampling series

There are connections here to the theory of single channel sampling (see, e.g., [3, Ch. 12]), but the present approach is much more general.
In order to match the operator reproducing equation (12) with a discrete analogue, some further assumption will have to be made. In fact we shall assume the existence of a sequence $(s_n) \subset E$, $n \in \mathbb{X}$ such that $\{\overline{h(s_n,t)}\}$ is an orthogonal basis for R with normalising factors ν_n, so that $\{\nu_n\overline{h(s_n,t)}\}$ is orthonormal. This can sometimes be traced back to the condition that $\{K_{s_n}\}$ be an orthogonal basis for $\mathcal{H}$. Again, we could assume that $\{h(s_n,t)\}$ is just a basis for R, or just a frame. However, weaker assumptions demand more technicalities and we will not pursue this kind of generality here.
Let $f \in R$. Its expansion in our assumed orthonormal basis is

$$f(t) = \sum_{n\in\mathbb{X}} c_n\nu_n\, \overline{h(s_n,t)} \quad (22)$$

where

$$c_n = \langle f, \nu_n\overline{h(s_n,\cdot)}\rangle = \overline{\nu_n}\langle f, h^*(\cdot,s_n\rangle = \overline{\nu_n}(\mathcal{B}^*f)(s_n)$$

by Theorem 5. So (22) is

$$f(t) = \sum_{n\in\mathbb{X}} |\nu_n|^2(\mathcal{B}^*f)(s_n)\, \overline{h(s_n,t)}. \quad (23)$$

Then (12) and (23) are concrete – discrete analogues of each other.

4. Multiplier operators with respect to the Fourier transform

Take E to be an LCA group G with dual Γ (for notations and references, see §2.2), and let R be a Paley–Wiener space PW_Ω. Let $\mu(\gamma)$ be a non-nul complex valued function on Γ such that

$$\begin{cases} 0 < \alpha \le |\mu(\gamma)| \le \beta < \infty, & \text{(Haar) a.a. } \gamma \in \Omega; \\ \mu(\gamma) = 0, & \gamma \notin \Omega. \end{cases} \quad (24)$$

Let $\mathcal{M}$ denote the operation of multiplication by $\chi_\Omega(\gamma)\mu(\gamma)$.

Definition 3 *Let $f \in PW_\Omega$. The operator $\mathcal{T}$ is defined by*

$$(\mathcal{T}f)(s) := (\mathcal{F}^{-1}\mathcal{M}\mathcal{F}f)(s)$$

Lemma 2 *The operator $\mathcal{T}$ of Definition 3 is a bijection on PW_Ω*

Proof Clearly $\mathcal{T}$ is linear. Furthermore it is one-to-one, since the null space of $\mathcal{T}$ is

$$\{f : \mathcal{T}f = \theta\} = \{f : \mu(\gamma)f^\wedge(\gamma) = \theta\}$$

which implies that $f = \theta$.
Again, $\mathcal{T}$ is "onto". Let $g \in PW_\Omega$. Then if $\mathcal{M}^{-1}$ denotes multiplication by $[\mu(\gamma)]^{-1}$, $f = \mathcal{F}^{-1}\mathcal{M}^{-1}\mathcal{F}g \in PW_\Omega$. Then from Definition 3, $\mathcal{T}f = g$.
The boundedness of $\mathcal{T}$ follows from two applications of Plancherel's Theorem. Indeed, let $f \in PW_\Omega$. Then

$$\begin{aligned}\|\mathcal{T}f\|_{L^2(G)} &= \|\mathcal{F}^{-1}\mathcal{M}\mathcal{F}f\|_{L^2(G)} = \|\mathcal{M}\mathcal{F}f\|_{L^2(\Gamma)} \\ &\le |\mu|\|\mathcal{F}f\|_{L^2(\Gamma)} = |\mu|\|f\|_{L^2(G)}.\end{aligned}$$

□

4.1 The operator kernel for $\mathcal{T}$

First we need to know the adjoint $\mathcal{T}^*$. Let f_1, $f_2 \in PW_\Omega$. The defining equation is

$$\langle \mathcal{T}f_1, f_2\rangle = \langle f_1, \mathcal{T}^* f_2\rangle.$$

Suppose that $\mathcal{T}^*$ is of the same form as $\mathcal{T}$ of Definition 3, that is,

$$\mathcal{T}^* f = \mathcal{F}^{-1}\mathcal{M}^*\mathcal{F}f, \tag{25}$$

where $\mathcal{M}^*$ denotes multiplication by the multiplier μ^* which is to be determined.
In the integral notation, and using the 'hat' notation for the Fourier transform, the criterion is:

$$\int_G \big(\mu(\cdot)f_1{}^\wedge(\cdot)\big)^\vee(t)\overline{f_2(t)}\,dt = \int_G f_1(t)\overline{\big(\mu^*(\cdot)f_2{}^\wedge(\cdot)\big)^\vee(t)}\,dt.$$

By Plancherel's theorem this is:

$$\int_\Gamma \mu(\gamma)f_1{}^\wedge(\gamma)\overline{f_2{}^\wedge(\gamma)}\,d\gamma = \int_\Gamma f_1{}^\wedge(\gamma)\overline{\mu^*(\gamma)f_2{}^\wedge(\gamma)}\,d\gamma,$$

from which we may choose $\mu^*(\gamma) = \overline{\mu(\gamma)}$.
It may be noted that $\mathcal{T}$ is self-adjoint if μ is real-valued.
It is now evident that the assumption (25) leads to

$$(\mathcal{T}^* f)(s) = \int_\Gamma \overline{\mu(\gamma)}f^\wedge(\gamma)(s,\gamma)\,d\gamma \tag{26}$$

The operator kernel for $\mathcal{T}$ can now be calculated. From Definition 1 and (7) we have

$$h(s,t) = \big(\mathcal{T}^*\varphi_\Omega(\cdot - t)\big)(s), \quad s,t \in G.$$

Therefore from (26), and using the 'shift' property of the Fourier transform,

$$\begin{aligned}h(s,t) &= \int_\Gamma \overline{\mu(\gamma)}\big(\chi_\Omega{}^\vee(\cdot - t)\big)^\wedge(\gamma)(s,\gamma)\,d\gamma \\ &= \int_\Gamma \overline{\mu(\gamma)}(-t,\gamma)\chi_\Omega(\gamma)(s,\gamma)\,d\gamma \\ &= \int_\Omega \overline{\mu(\gamma)}(s-t,\gamma)\,d\gamma \\ &= \overline{\mu(\cdot)}^\vee(s-t).\end{aligned}$$

Hence

$$\overline{h(s,t)} = \overline{\overline{\mu(\cdot)}^\vee(s-t)} = \mu^\wedge(s-t).$$

Now (12) becomes

$$f(t) = \langle (\mathcal{T}^{-1}f)(\cdot), \overline{\mu}^\vee(\cdot - t)\rangle, \tag{27}$$

and (23) becomes

$$f(t) = \sum_{n\in\mathbb{X}} |\nu_n|^2 (\mathcal{T}^* f)(s_n)\,\mu^\wedge(s_n - t). \tag{28}$$

5. Examples

Example 1 *The classical case*

Naturally, we expect to recover the case of the classical reproducing equation and sampling formula as special cases of the theory. To do this we pick $G = \mathbb{R}$, $\Omega = [-\pi,\pi]$, $\mathcal{T} = \mathcal{I} = \mathcal{T}^* = \mathcal{T}^{-1}$ and $\mu = \chi_{[-\pi,\pi]}(y)$. Therefore we have

$$\mu^\vee(s-t) = \frac{1}{\sqrt{2\pi}}\int_{-\pi}^{\pi} e^{i(s-t)y}\,dy = \sqrt{2\pi}\,\mathrm{sinc}(s-t).$$

Here and in subsequent Examples the choice of Haar measure on G, Γ, etc., accounts for apparent anomalies in the normalising constants in the formulae (e.g., Haar measure on $\mathbb{R}$ is taken to be $(2\pi)^{-1/2}$ times Lebesgue measure. See [2, p. 257]). With these choices, (27) becomes (1).
The classical sampling series (2) now follows the textbook proof. Since $\{e^{-iny}/\sqrt{2\pi} : n \in \mathbb{Z}\}$ is an orthonormal (ON) basis of $L^2(-\pi,\pi)$, Plancherel's theorem shows that the inverse Fourier transforms $\{\mathrm{sinc}(n-t)\} : n \in \mathbb{Z}\}$ form an orthonormal basis of $PW_{[-\pi,\pi]}$. Coefficients in the expansion of f in this basis are obtained from (1) and so, with $s_n = n$, our choices for $\mathcal{T}$ and μ show that (28) becomes (2).

Example 2 *The Hilbert transform*

Another well-known example illustrates the present theory; a member of $PW_{[-\pi,\pi]}$ can be sampled and reconstructed from samples of its Hilbert transform (see, e.g., [3, p. 126] and references there). This idea can be fitted it into the theme of the present lecture by taking $G = \mathbb{R}$, $\Omega = [-\pi,\pi]$, $\mathcal{T} = \mathcal{H} := \mathcal{F}^{-1}\mathcal{M}\mathcal{F}$ where $\mathcal{M}$ denotes multiplication by $-i\,\mathrm{sgn}(y)$. $\mathcal{H}$ is the Hilbert transform on $PW_{[-\pi,\pi]}$.

For (27) we need

$$\begin{aligned}\overline{\mu}^{\vee}(s-t) &= \frac{1}{\sqrt{2\pi}}\frac{i}{\sqrt{2\pi}}\int_{-\pi}^{\pi}\operatorname{sgn}(y)e^{i(s-t)y}\,dy \\ &= -\operatorname{sinc}\tfrac{1}{2}(s-t)\sin\tfrac{\pi}{2}(s-t) \qquad (29)\end{aligned}$$

after a simple calculation. Also we have $\mathcal{H}^{-1} = -\mathcal{H} = \mathcal{H}^*$, therefore (27) is

$$f(t) = -\int_{\mathbb{R}}\big(\mathcal{H}f\big)(\tau)\operatorname{sinc}\tfrac{1}{2}(\tau-t)\sin\tfrac{\pi}{2}(\tau-t)\,d\tau. \quad (30)$$

For (28) we need to find $\{s_n\}$ such that $\{\mu^{\wedge}(s_n-t)\}$, $n \in \mathbb{Z}$, is an ON basis of PW_{π}. We can start with the ON basis $\{e^{iny}/\sqrt{2\pi}\}$, $(n \in \mathbb{Z})$, of $L^2(-\pi,\pi)$, then multiply each member by $-i\operatorname{sgn}(y)$. The result is again an ON basis, as a consequence of $|-i\operatorname{sgn}(y)| = 1$ a.e. on $[-\pi,\pi]$. The inverse Fourier transform of a typical one of these basis elements is

$$\frac{-i}{2\pi}\,\mathcal{F}^{-1}\big(\operatorname{sgn}(\cdot)e^{-in\cdot}\big)(t) = -\operatorname{sinc}\tfrac{1}{2}(n-t)\sin\tfrac{\pi}{2}(n-t)$$

by the same calculation as in (29). But, taking account of Haar measure, this also gives $\mu^{\wedge}(n-t)$. Hence (28) becomes

$$f(t) = \sum_{n\in\mathbb{Z}}\big(\mathcal{H}f\big)(n)\operatorname{sinc}\tfrac{1}{2}(n-t)\sin\tfrac{\pi}{2}(n-t) \quad (31)$$

Our concrete – discrete comparison is between (30) and (31).

Example 3 *The Riesz transforms*

For background on the Riesz transforms see [9, p. 223] Take G to be $\mathbb{R}^d$, $(d \in \mathbb{N})$. Let $\mathbf{t} = (t_1,\dots,t_d)$ and let $\boldsymbol{y} = (y_1,\dots,y_d)$ etc. Let the scalar product in $\mathbb{R}^d$ be denoted by $\langle\ ,\rangle$.

Definition 4 *Let $f \in L^2(\mathbb{R}^d)$, and define*

$$\mathcal{R}_j f := \mathcal{F}^{-1}\mathcal{M}_j\mathcal{F}f, \quad j = 1,\dots,d, \qquad (32)$$

$\mathcal{M}_j$ denoting multiplication by $-iy_j/|\boldsymbol{y}|\ \chi_{[-\pi,\pi]^d}(\boldsymbol{y})$.

We note that this multiplier is not bounded away from zero when $d \geq 2$ and $\boldsymbol{y} \in [-\pi,\pi]^d$ and therefore does not always satisfy the criterion (24). However, it is possible to define operators involving the Riesz transforms which do satisfy the criterion (24).
First we consider the case $d = 2$, and define the operator

$$\mathcal{R} := \mathcal{R}_1 + i\mathcal{R}_2 \qquad (33)$$

acting on $PW_{[-\pi,\pi]^2}$. Its multiplier is

$$m(\boldsymbol{y}) := (-i)\left(\frac{y_1}{|\boldsymbol{y}|} + i\frac{y_2}{|\boldsymbol{y}|}\right)$$

and clearly we have $|m(\boldsymbol{y})| = 1$ a.e. Hence m satisfies the criterion (24) with respect to two-dimensional Lebesgue measure. Now (27) becomes

$$f(\mathbf{t}) = \frac{1}{2\pi}\int_{\mathbb{R}^2}\big(\mathcal{R}^{-1}f\big)(\mathbf{s})\overline{m}^{\vee}(\mathbf{s}-\mathbf{t})\,d\mathbf{s}. \qquad (34)$$

Since the multiplier is of unit modulus, a two dimensional version of the construction that we used in the previous example shows that

$$\left\{\frac{-i}{2\pi}\left(\frac{y_1}{|\boldsymbol{y}|} + i\frac{y_2}{|\boldsymbol{y}|}\right)e^{-i\langle\mathbf{k},\boldsymbol{y}\rangle}\right\}, \quad (\mathbf{k}\in\mathbb{Z}^2),$$

is an ON basis of $L^2([-\pi,\pi]^2)$. Then (28) becomes

$$f(\mathbf{t}) = \sum_{\mathbf{k}\in\mathbb{Z}^2}\big(R^*f\big)(\mathbf{k})m^{\wedge}(\mathbf{k}-\mathbf{t}). \qquad (35)$$

The comparison for this example lies between the concrete (34) and the discrete (35).
Other combinations of the Riesz transforms are possible, in two and higher dimensions, whose multipliers satisfy (24) but are not always of unit modulus.

6. Conclusions

The multiplier transforms treated in this study form a rather restricted class of operators; nevertheless, the methods can be used in connection with the very important Riesz transforms. It remains to investigate extensions to other types of operator. Likely candidates are, for example, multiplier transforms with less restrictive conditions on the multiplier, the singular integral operators of Calderón–Zygmund type (a class containing the Riesz transforms, see, e.g., [9, Ch. VI]), and operators of the Hankel and Toeplitz type (e.g., [7]).

References:

[1] M.M. Dodson. Groups and the sampling theorem. *Sampl. Theory Signal Image Process.*, 6(1):1–27, 2007.

[2] M.M. Dodson and M.G. Beaty. Abstract harmonic analysis and the sampling theorem. In J.R. Higgins and R.L. Stens, editors, *Sampling theory in Fourier and signal analysis: advanced topics*, pages 233–265. Clarendon Press, Oxford, 1999.

[3] J.R. Higgins. *Sampling theory in Fourier and signal analysis: foundations*. Clarendon Press, Oxford, 1996.

[4] I. Kluvánek. Sampling theorem in abstract harmonic analysis. *Mat.-Fyz. Casopis Sloven. Akad. Vied.*, 15:43–48, 1965.

[5] H. Meschkowski. *Hilbertsche Räume mit Kernfunktion*. Springer–Verlag, Berlin, 1962.

[6] F. Riesz and B. Sz.-Nagy. *Functional analysis*. Dover Publications, New York, 1990.

[7] R. Rochberg. Toeplitz and Hankel operators on the Paley–Wiener space. *Integral Equations Operator Theory*, 10(2), 1987.

[8] S. Saitoh. *Integral transforms, reproducing kernels and their applications*. Longman, Harlow, 1997.

[9] E.M. Stein and G. Weiss. *Introduction to Fourier analysis on Euclidean spaces*. Princeton University Press, Princeton, 1971.

Explicit localization estimates for spline-type spaces

José Luis Romero

Departamento de Matemática
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria, Pabellón I
1428 Capital Federal
ARGENTINA
and CONICET, Argentina.
jlromero@dm.uba.ar

Abstract:

We give some explicit decay estimates for the dual system of a basis of functions that are polynomially localized in space.

1. Introduction

A spline-type space S is a closed subspace of $L^2(\mathbb{R}^d)$ possessing a Riesz basis of functions well localized is space. That is, there exists a family of functions $\{f_k\}_k \subseteq S$ and constants $0 < A \leq B < +\infty$ such that

$$A\|c\|_{\ell^2} \leq \|\sum_k c_k f_k\|_{L^2} \leq B\|c\|_{\ell^2}, \qquad (1)$$

holds for every $c \in \ell^2$, and the functions $\{f_k\}_k$ satisfy an spatial localization condition.

In a spline-type space any function in $f \in S$ has a unique expansion $f = \sum_k c_k f_k$. Moreover the coefficients are given by $c_k = \langle f, g_k \rangle$, where $\{g_k\}_k \subseteq S$ is the *dual basis*, a set of functions characterized by the relation $\langle g_k, f_j \rangle = \delta_{k,j}$. These spaces provide a very natural framework for the sampling problem.

The general theory of localized frames (see [6], [5] and [2]) asserts that the functions forming the dual basis satisfy a similar spatial localization. This can be used to extend the expansion in (1) to other spaces, so that the family $\{f_k\}_k$ becomes a Banach frame for an associated family of Banach spaces (see [4] and [6]). In the case of a spline-type space S, this means that the decay of a function in S can be characterized by the decay of its coefficients and, in particular, that the functions $\{f_k\}_k$ form a so called p-Riesz basis for its L^p-closed linear span, for the whole range $1 \leq p \leq \infty$.

We derive, in some concrete case, explicit bounds for the localization of the dual basis. We will work with a set of functions satisfying a polynomial decay condition around a set of nodes forming a lattice. By a change of variables, we can assume that the lattice is $\mathbb{Z}^d$. So, we will consider a set of functions $\{f_k\}_k \subseteq L^2(\mathbb{R}^d)$ satisfying the condition,

$$|f_k(x)| \leq C\,(1 + |x - k|)^{-s}, \quad x \in \mathbb{R}^d \text{ and } k \in \mathbb{Z}^d,$$

for some constant C. This type of spatial localization is specifically covered by the results in [5], but the constants given there are not explicit. We will derive a polynomial decay condition for the dual basis $\{g_k\}_k$, giving explicit information on the resulting constants. This yields some qualitative information, like the dependence of theses constants on A, C and s and the corresponding p-Riesz basis bounds for the original basis.

2. Main result

Theorem 1 *Let $C \geq 1$, and let $t > d$ be integers. Let $s > d + t$ be a real number. For $k \in \mathbb{Z}^d$ let $f_k : \mathbb{R}^d \to \mathbb{C}$ be a measurable function such that*

$$|f_k(x)| \leq C\,(1 + |x - k|)^{-s}, \qquad (x \in \mathbb{R}^d).$$

Suppose that $\{f_k\}_k$ is a Riesz basis for its L^2 closed linear span S, with bounds $0 < A \leq B < \infty$. Let $\{g_k\}_k \subseteq S$ be its dual basis.

Then, the dual functions satisfy,

$$|g_k(x)| \leq D\,(1 + |x - k|)^{-t}, \qquad (x \in \mathbb{R}^d).$$

where D is given by,

$$D = \frac{E^{st} C^{2t+1}}{(s - t - d)^t} \frac{1 + A^{t-1}}{A^{t+1}},$$

for some constant $E > 0$ that only depends on the dimension d.

Remark 1 *The constant E can be explicitly determined from the proof.*

The results in [6] prescribe polynomial decay estimates for the dual basis similar to those possessed by the original basis. As a trade-off for the explicit constants we will not obtain the full preservation of these decay conditions. Nevertheless, any degree of polynomial decay on the dual system can be granted, provided that the original basis has sufficiently good decay.

Finally observe that, although the basis $\{f_k\}_k$ is assumed to be concentrated around a lattice of nodes, the functions f_k are not assumed to be shifts of a single function. In particular, Theorem 1 below allows for a basis of functions whose 'optimal' concentration nodes do not form a lattice but are comparable to one. The 'eccentricity' of the configuration of concentration nodes is, however, penalized by the constants modelling the decay.

3. Sketch of a proof and comments

Now we sketch the proof of the main result, for a complete proof see [11].
Consider the gram matrix of the basis $\{f_k\}_k$ given by,

$$M \equiv (m_{k,j})_{k,j\in\mathbb{Z}^d}, \quad m_{k,j} := \langle f_k, f_j \rangle .$$

Since $\{f_k\}_k$ is a Riesz sequence, M, as an operator on ℓ^2, has an inverse $N \equiv (n_{k,j})_{k,j\in\mathbb{Z}^d}$. Moreover, $\|N\|_{\ell^2\to\ell^2} \leq A^{-1}$ and $n_{k,j} = \langle g_k, g_j \rangle$, where $\{g_k\}_k \subseteq S$ is the dual basis of $\{f_k\}_k$.
The localization assumptions on the basis $\{f_k\}_k$ yield a polynomial decay estimate on the entries of M,

$$|m_{k,j}| \lesssim (1 + |k - j|)^{-s}.$$

If we can establish a similar estimate for the entries of N,

$$|n_{k,j}| \lesssim (1 + |k - j|)^{-t}.$$

with all the constants given explicitly, then, using calculations similar to those in [5], we obtain the desired polynomial concentration conditions for the dual functions.
Let us first consider the case where the basis $\{f_k\}_k$ consists of integer shifts of a single generator f (that is, $f_k = f(\cdot - k), k \in \mathbb{Z}^d$). In this case, the matrix M is constant on its diagonals. That is,

$$m_{k,j} = a_{k-j},$$

for some sequence a. Similarly, N is given by

$$n_{k,j} = b_{k-j},$$

where the sequence b satisfies $a * b = \delta$.
Therefore, in this special case, M and N are convolution operators. The off-diagonal decay of their entries is equivalent to the decay of their kernels a and b. Since the decay of a sequence x can be characterized by the smoothness of its Fourier transform $\hat{x}$, the problem can be reformulated as the preservation of the smoothness of the function $\hat{a}$ under pointwise inversion. This reasoning is present, for example, in [1].
We can measure the smoothness of $\hat{a}$ by considering weak-derivatives and use repeatedly a chain-rule argument for Sobolev spaces to obtain similar smoothness conditions for $\hat{b}$.
In the general case, where M and N need not be convolution operators, we try to imitate this reasoning, but we avoid using the Fourier transform.
Given a matrix $L \equiv (l_{k,j})_{k,j\in\mathbb{Z}^d}$ and $1 \leq h \leq d$, we consider the new matrix,

$$D_h(L)_{k,j} := (k_h - j_h) l_{k,j}.$$

Observe that, up to some multiplicative constant, the map D_h acts on a convolution operator by taking a partial derivative of its symbol (that is, the Fourier transform of its kernel.) The domain of D_h consists of those matrices L such that $D_h(L)$ defines a bounded operator on ℓ^2. We call $D_h(L)$ the *derivative* of L (with respect to x_h.)
D_h is a derivation in the sense that it satisfies the equation $D_h(AB) = D_h(A)B + AD_h(B)$, provided that $D_h(A)$ and $D_h(B)$ are both defined. Derivations are a well-known tool in operator-algebras theory (see [3], [9] and [10].)
Since $MN = I$ and $D_h(I) = 0$, we can formally express the high-order derivatives of N in terms of its lower-order ones and all the derivatives of M,

$$D_h^u(N) = -\sum_{l=0}^{u-1} \binom{u}{l} D_h^l(N) D_h^{u-l}(M) N. \qquad (2)$$

Using the polynomial off-diagonal decay bounds on M and the bound $\|N\|_{\ell^2\to\ell^2} \leq A^{-1}$ we can obtain bounds for the $\ell^2 \to \ell^2$ norms of some derivatives of N. These imply polynomial off-diagonal decay estimates for N, and hence yield the desired spatial localization bounds for the dual basis.
In the argument above we related the off-diagonal decay of a matrix with the $\ell^2 \to \ell^2$ norm of its derivatives. The $\ell^2 \to \ell^2$ norm of a matrix is not determined by the size of its entries. However, there are some necessary and (other) sufficient conditions on the size of the entries of a matrix for it to be bounded on ℓ^2. This "gap" in the conditions accounts for the loss of some decay information in Theorem 1, when passing from the original basis to its dual system.
Finally we point out that the formal computations in the above argument are not sufficient to prove the theorem. Consider again the simple case of a basis of integer shifts. With the notation of the discussion above, we have the relation

$$a * b = \delta, \qquad (3)$$

we have some decay estimate on a (that can be reformulated as a smoothness condition on $\hat{a}$) and we want to prove a similar decay condition for b. There may be various sequences x satisfying the relation $a * x = \delta$; b can be singled out as the only one of them having a bounded Fourier transform. For example, when a is finitely supported, equation 3 is a linear difference equation which has other solutions besides b (that grow exponentially). The decay of the sequence b can be rigorously proved by resorting to some Sobolev-space smoothing argument.
In the general case, to derive equation (2), one needs to use the associativity of the product of matrices. This is justified only if all the matrices involved represent bounded operators. In other words, we need to know a priori that the derivatives of N that are involved in equation (2) define bounded operators. This can be proved using the general results on derivations on Banach algebras (see [3], [9]) or Jaffard's Theorem [8].
The use of derivations is somehow implicit in Jaffard's paper [8]. Recently, Gröchenig and Klotz [7] have systematically studied the use of derivations in connection to various problems including the preservation under inversion of various kinds of off-diagonal decay conditions.

4. Application

From Theorem 1 we can derive the following qualitative statement.

Theorem 2 *Let $\{F^i\}_{i\in I}$ be a family of Riesz sequences,*

$$F^i \equiv \{f_k^i\}_{k\in\mathbb{Z}^d} \subseteq \boldsymbol{L}^2(\mathbb{R}^d), \quad (i \in I).$$

sharing a uniform lower basis bound. Suppose that the family $\left\{F^i\right\}_i$ *satisfies a uniform concentration condition,*

$$\left|f_k^i(x)\right| \leq C\left(1+|x-k|\right)^{-s}, (x \in \mathbb{R}^d, k \in \mathbb{Z}^d, i \in I),$$

for some constants $C \geq 1$, $s > d+t$ *and* $t > d$, *with* t *an integer.*
Then the following holds.

(a) *The respective family of dual systems* $\left\{G^i\right\}_i$ *- where* $G^i \equiv \left\{g_k^i\right\}_{k\in\mathbb{Z}^d}$ *- satisfies a uniform concentration condition,*

$$\left|g_k^i(x)\right| \leq D\left(1+|x-k|\right)^{-t}, (x \in \mathbb{R}^d, k \in \mathbb{Z}^d, i \in I),$$

for some constant $D \geq 1$.

(b) *A uniform* p*-Riesz basis condition holds, for all* $1 \leq p \leq \infty$. *More precisely, there exist constants* $q, Q > 0$ *such that for any* $p \in [1,\infty]$ *and any* $i \in I$, *the relation*

$$q\|c\|_{\ell^p} \leq \|\sum_k c_k f_k^i\|_{L^p} \leq Q\|c\|_{\ell^p}$$

holds for all finitely supported sequences $(c_k)_{k\in\mathbb{Z}^d}$.

Statement (a) follows directly from Theorem 1. Examining the proofs in [5] we see that the uniformity of the constants given in (a) yields statement (b).
This qualitative conclusion on Theorem 2 was the original motivation for Theorem 1.
Finally, observe that the arguments given above are applicable to a generel *intrinsically localized* basis in the sense of [5].

5. Acknowledgements

The author wishes to thank Karlheinz Gröchenig and Andreas Klotz for their comments and for sharing an early draft of [7], and is indebted to Hans Feichtinger and Ursula Molter for some insightful discussions.
The author holds a fellowship from the CONICET and thanks this institution for its support. His research is also partially supported by grants: PICT06-00177, CONICET PIP N 5650, UBACyT X149.
This note was partially written during a long-term visit to NuHAG in which the author was supported by the EUCETIFA Marie Curie Excellence Grant (FP6-517154, 2005-2009).

References:

[1] Akram Aldroubi and Karlheinz Gröchenig. Nonuniform sampling and reconstruction in shift-invariant spaces. *SIAM Rev.*, 43(4):585–620, 2001.

[2] Radu M. Balan, Peter G. Casazza, Christopher Heil, and Z. Landau. Density, overcompleteness, and localization of frames I: Theory. *J. Fourier Anal. Appl.*, 12(2):105–143, 2006.

[3] Ola Bratteli and Derek W. Robinson. Unbounded derivations of c*-algebras. *Commun. math. Phys*, 42:253–268, 1975.

[4] Hans G. Feichtinger and Karlheinz Gröchenig. Banach spaces related to integrable group representations and their atomic decompositions, I. *J. Funct. Anal.*, 86:307–340, 1989. reprinted in 'Fundamental Papers in Wavelet Theory' Heil, Christopher and Walnut, David F.(2006).

[5] Massimo Fornasier and Karlheinz Gröchenig. Intrinsic localization of frames. *Constr. Approx.*, 22(3):395–415, 2005.

[6] Karlheinz Gröchenig. Localization of Frames, Banach Frames, and the Invertibility of the Frame Operator. *J. Fourier Anal. Appl.*, 10(2):105–132, 2004.

[7] Karlheinz Gröchenig and Klotz Andreas. Noncommutative approximation: Inverse-closed subalgebras and off-diagonal decay of matrices. *Preprint*, available at http://arxiv.org/abs/0904.0386, 2009.

[8] Stephane Jaffard. Propriétés des matrices "bien localisées" près de leur diagonale et quelques applications. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 7(5):461–476, 1990.

[9] Edward Kissin and Victor Shulman. Dense q-subalgebras of banach and c*-algebras and unbounded derivations of banach and c*-algebras. *Proc. Edinburgh Math. Soc*, 36:261–276, 1993.

[10] Edward Kissin and Victor Shulman. Differential properties of some dense subalgebras of c*-algebras. *Proc. Edinburgh Math. Soc*, 37:399–422, 1994.

[11] José Luis Romero. Explicit localization estimates for spline-type spaces. *Submitted*, available at http://arxiv.org/abs/0902.0557, 2008.

A *Fast* Fourier Transform with Rectangular Output on the BCC and FCC Lattices

Usman R. Alim (1) and Torsten Möller (1)

(1) School of Computing Science, Simon Fraser University, Burnaby BC V5A 1S6, Canada.
ualim@cs.sfu.ca, torsten@cs.sfu.ca

Abstract:

This paper discusses the efficient, non-redundant evaluation of a Discrete Fourier Transform on the three dimensional Body-Centered and Face-Centered Cubic lattices. The key idea is to use an axis aligned window to truncate and periodize the sampled function which leads to separable transforms. We exploit the geometry of these lattices and show that by choosing a suitable non-redundant rectangular region in the frequency domain, the transforms can be efficiently evaluated using the Fast Fourier Transform.

1. Introduction

The Discrete Fourier Transform (DFT) is an important tool used to analyze and process data in an arbitrary number of dimensions. Most applications of the DFT in higher dimensions, however, rely on a tensor product extension of a one-dimensional DFT, with the assumption that the underlying data is sampled on a Cartesian lattice. This extension has the advantage that it allows for a straightforward application of the Fast Fourier Transform (FFT).

The Cartesian lattice is known to be sub-optimal when it comes to sampling a band-limited function in two or higher dimensions [6]. In 3D, for instance, the Body-Centered Cubic (BCC) lattice is the optimal sampling lattice and yields a 30% savings in samples as compared to the Cartesian lattice [8]. The Face-Centered Cubic (FCC) lattice, although not optimal, is still better than the Cartesian lattice and is also the lattice that yields the minimum amount of Fourier-domain aliasing when sampling a general trivariate function [2].

From the perspective of continuous signal reconstruction, both the BCC and FCC lattices have received considerable attention because of their many applications in Visualization and Computer Graphics. Entezari et al. have devised a set of Box-Splines that can be used for signal approximation on the BCC [3] as well as the FCC [5] lattices. However, very little effort has gone into the development of discrete processing tools that are suitable for these non-Cartesian lattices.

The idea of a multidimensional DFT (MDFT) on non-Cartesian lattices is not new. Mersereau provided a derivation of a DFT for a hexagonally periodic sequence and designed other digital filters suitable for a 2D hexagonal lattice [6]. Later, the idea was extended to higher dimensions and a MDFT for arbitrary sampling lattices was proposed [7]. Guessoum et al. proposed an algorithm for evaluating the MDFT that has the same computational complexity as the Cartesian DFT [4].

Recently, Csébfalvi et al. [1] applied the MDFT to the BCC and FCC lattices by choosing a Cartesian periodicity in the spatial domain which leads to a Cartesian sampling of the Fourier transform. This allows the MDFT to be written in a separable form that can be evaluated via the FFT. However, their representation is redundant and leads to inefficient transforms. The aim of this paper is to revisit these transforms and show that they can be computed much more efficiently by exploiting the geometric properties of the BCC and FCC lattices to eliminate the redundancy.

The paper is organized as follows. We provide a basic review of multidimensional sampling in Section 2. which is later used in the derivation of a fast DFT for BCC and FCC lattices in Section 3. Some properties of these transforms are discussed in Section 4. and a summary is presented in Section 5.

2. Optimal Trivariate Sampling

Let $f_c(\boldsymbol{x})$ be a continuous trivariate function and $F_c(\boldsymbol{\xi})$ be its Fourier transform defined as

$$F_c(\boldsymbol{\xi}) = \int_{\mathbb{R}^3} f_c(\boldsymbol{x}) \exp[-2\pi j \boldsymbol{\xi}^T \boldsymbol{x}] d\boldsymbol{x} \qquad (1)$$

where T denotes the transpose operation. Let $f(\boldsymbol{n})$ be the sampled sequence obtained by sampling the function through

$$f(\boldsymbol{n}) = f_c(\boldsymbol{L}\boldsymbol{n}) \qquad (2)$$

where $\boldsymbol{L}$ is a 3×3 sampling matrix and $\boldsymbol{n}$ is an integer vector. Sampling on the lattice defined by the matrix $\boldsymbol{L}$ amounts to a periodization of the Fourier spectrum on a reciprocal lattice generated by the matrix $\boldsymbol{L}^{-T}$. In particular, the spectrum of the sampled sequence is given by [9]

$$\hat{F}(\boldsymbol{\xi}) = \frac{1}{|\det \boldsymbol{L}|} \sum_{\boldsymbol{r}} F_c(\boldsymbol{\xi} - \boldsymbol{L}^{-T}\boldsymbol{r}) \qquad (3)$$

where $\boldsymbol{r}$ is any integer vector.

If we assume that $f_c(\boldsymbol{x})$ is isotropically band-limited (i.e $F_c(\boldsymbol{\xi}) = 0$ for $\|\boldsymbol{\xi}\| > \xi_0$ for some band-limit ξ_0), then one of the lattices that achieves the tightest possible packing of the spectrum replicas (spheres) in the Fourier domain is

the FCC lattice. Thus, in order to sample a trivariate band-limited function optimally, the function should be sampled on the reciprocal of the FCC lattice, i.e. the BCC lattice.

3. Discrete Fourier Transform

If the sequence $f(\boldsymbol{n})$ is non-zero within a finite region, it can be periodically extended spatially and represented as a Fourier series which is a sampled version of the transform (3) [7]. The pattern with which the continuous transform (3) is sampled in the Fourier domain depends on the periodicity pattern in the spatial domain. Merserau et al. [7] used a periodicity matrix to define the periodic extension of the finite sequence. Here, we use a somewhat different approach by splitting the sampled sequence into constituent Cartesian sequences [1].

The BCC and FCC lattices $\mathcal{L}_{\mathsf{B}}$ and $\mathcal{L}_{\mathsf{F}}$ are generated by the integer sampling matrices

$$\boldsymbol{L}_{\mathsf{B}} = \begin{bmatrix} 1 & -1 & 1 \\ -1 & 1 & 1 \\ 1 & 1 & -1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{L}_{\mathsf{F}} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

respectively. Both these lattices are based on a cubic sampling pattern whereby, in addition to samples at the eight corners of a cube, $\mathcal{L}_{\mathsf{B}}$ has an additional sample in the center of the cube and $\mathcal{L}_{\mathsf{F}}$ has six additional samples on the faces. Both these lattices can also be built from shifts of a Cartesian sublattice as shown in Fig. 1. In particular, samples that lie on the corners of cubes form the sublattice $2\mathbb{Z}^3$. The quotient group $\mathcal{L}_{\mathsf{B}}/2\mathbb{Z}^3$ is isomorphic to $\mathbb{Z}_2$ and the quotient group $\mathcal{L}_{\mathsf{F}}/2\mathbb{Z}^3$ is isomorphic to $\mathbb{Z}_4$. Therefore, $\mathcal{L}_{\mathsf{B}}$ can be partitioned into two Cartesian cosets while $\mathcal{L}_{\mathsf{F}}$ has four Cartesian cosets (Fig. 1).

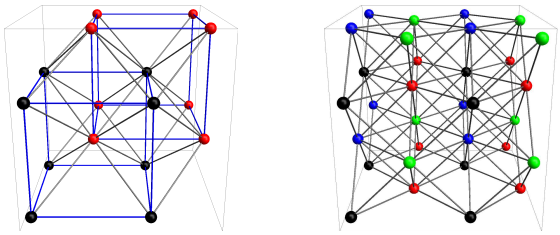


Figure 1: Left, the BCC lattice, a 16 point view. Right, the FCC lattice, a 32 point view. Lattice sites that are Voronoi neighbors are linked to each other. Cosets are indicated by different colors.

3.1 BCC DFT

The BCC lattice with arbitrary scaling is obtained via the sampling matrix $h\boldsymbol{L}_{\mathsf{B}}$ where h is a positive scaling parameter. The Voronoi cell is a truncated octahedron having a volume of $|\det h\boldsymbol{L}_{\mathsf{B}}| = 4h^3$. The Voronoi cell of the reciprocal FCC lattice is a rhombic dodecahedron having a volume of $\frac{1}{4h^3}$. Since $\mathcal{L}_{\mathsf{B}}$ has two Cartesian cosets, a sampled sequence can be split up into two subsequences given by

$$f_0(\boldsymbol{n}) = f_c(2h\boldsymbol{I}\boldsymbol{n}) \quad \text{and} \quad f_1(\boldsymbol{n}) = f_c(2h\boldsymbol{I}\boldsymbol{n} + h\boldsymbol{t}),$$

where $\boldsymbol{I}$ is the 3×3 identity matrix, $\boldsymbol{t}$ is the translation vector $(1,1,1)^T$ and $\boldsymbol{n} = (n_1, n_2, n_3)^T$ is any integer vector. $f_0(\boldsymbol{n})$ is the sequence associated with the first coset while $f_1(\boldsymbol{n})$ is associated with the second. Since these sequences are sampled on a Cartesian pattern, a straightforward truncation of the original sequence is to choose a cuboid shaped fundamental region generated by limiting $\boldsymbol{n}$ to the set $\mathcal{N} := \{\boldsymbol{n} \in \mathbb{Z}^3 : 0 \le n_1 < N_1, 0 \le n_2 < N_2, 0 \le n_3 < N_3\}$ for some positive integers N_1, N_2 and N_3. This region consists of $2N_1N_2N_3$ data points (i.e. Voronoi cells) and has a total volume of $8N_1N_2N_3h^3$. If we define $\boldsymbol{N}$ to be the diagonal matrix $\mathrm{diag}(N_1, N_2, N_3)$, then the two subsequences $f_0(\boldsymbol{n})$ and $f_1(\boldsymbol{n})$ contained within the fundamental region can be periodically extended on a Cartesian pattern such that they satisfy

$$f_0(\boldsymbol{n} + \boldsymbol{N}\boldsymbol{r}) = f_0(\boldsymbol{n}) \quad \text{and} \quad f_1(\boldsymbol{n} + \boldsymbol{N}\boldsymbol{r}) = f_1(\boldsymbol{n}),$$

for all $\boldsymbol{n}$ and $\boldsymbol{r}$ in $\mathbb{Z}^3$.

This Cartesian periodic extension in the spatial domain amounts to a Cartesian sampling in the Fourier domain. In particular, the continuous transform (3) is sampled at the frequencies $\boldsymbol{\xi} = \frac{1}{2h}\boldsymbol{N}^{-1}\boldsymbol{k}$ yielding the sequence

$$\begin{aligned} F(\boldsymbol{k}) =& \hat{F}(\boldsymbol{\xi})\big|_{\boldsymbol{\xi}=\frac{1}{2h}\boldsymbol{N}^{-1}\boldsymbol{k}} \\ =& \sum_{\boldsymbol{n}\in\mathcal{N}} f_0(\boldsymbol{n}) \exp\big[\frac{-2\pi j}{2h}\boldsymbol{k}^T\boldsymbol{N}^{-1}2h\boldsymbol{I}\boldsymbol{n}\big] + \\ & f_1(\boldsymbol{n}) \exp\big[\frac{-2\pi j}{2h}\boldsymbol{k}^T\boldsymbol{N}^{-1}(2h\boldsymbol{I}\boldsymbol{n} + h\boldsymbol{t})\big] \\ =& \sum_{\boldsymbol{n}\in\mathcal{N}} \big(f_0(\boldsymbol{n}) + f_1(\boldsymbol{n}) \exp\big[-\pi j\boldsymbol{k}^T\boldsymbol{N}^{-1}\boldsymbol{t}\big]\big) \cdot \\ & \exp\big[-2\pi j\boldsymbol{k}^T\boldsymbol{N}^{-1}\boldsymbol{n}\big], \end{aligned} \tag{4}$$

where $\boldsymbol{k} = (k_1, k_2, k_3)^T \in \mathbb{Z}^3$ is the frequency index vector. The above equation defines a forward BCC DFT. Since it is a sampled version of a continuous transform that is periodic on an FCC lattice, it should be invariant under translations that lie on the reciprocal lattice generated by the matrix $(h\boldsymbol{L}_{\mathsf{B}})^{-T} = \frac{1}{2h}\boldsymbol{L}_{\mathsf{F}}$. This property is easily demonstrated as follows. If $\boldsymbol{r} \in \mathbb{Z}^3$, then after substituting $\boldsymbol{\xi} = \frac{1}{2h}(\boldsymbol{N}^{-1}\boldsymbol{k} + \boldsymbol{L}_{\mathsf{F}}\boldsymbol{r})$ in (4) and simplifying, we get

$$\begin{aligned} & \hat{F}\big(\frac{1}{2h}(\boldsymbol{N}^{-1}\boldsymbol{k} + \boldsymbol{L}_{\mathsf{F}}\boldsymbol{r})\big) \\ =& \sum_{\boldsymbol{n}\in\mathcal{N}} \big(f_0(\boldsymbol{n}) + f_1(\boldsymbol{n}) \exp\big[-\pi j(\boldsymbol{k}^T\boldsymbol{N}^{-1} + \boldsymbol{r}^T\boldsymbol{L}_{\mathsf{F}})\boldsymbol{t}\big]\big) \cdot \\ & \exp\big[-2\pi j(\boldsymbol{k}^T\boldsymbol{N}^{-1} + \boldsymbol{r}^T\boldsymbol{L}_{\mathsf{F}})\boldsymbol{n}\big] \\ =& \hat{F}(\frac{1}{2h}\boldsymbol{N}^{-1}\boldsymbol{k}), \end{aligned}$$

since $\boldsymbol{r}^T\boldsymbol{L}_{\mathsf{F}}\boldsymbol{n}$ is always an integer and $\boldsymbol{r}^T\boldsymbol{L}_{\mathsf{F}}\boldsymbol{t}$ is always even.

One fundamental period of the BCC DFT is contained within a rhombic dodecahedron of volume $\frac{1}{4h^3}$. The sampling density in the frequency domain is given by $|\det \frac{1}{2h}\boldsymbol{N}^{-1}| = (8N_1N_2N_3h^3)^{-1}$. Thus, the fundamental period consists of a total of $2N_1N_2N_3$ distinct frequency samples which is the same as the number of distinct spatial samples.

The inverse BCC DFT is obtained by summing over all the distinct sinusoids and evaluating them at the spatial sample locations. This gives

$$f_0(\boldsymbol{n}) = \frac{1}{N}\sum_{\boldsymbol{k}\in\mathcal{K}} F(\boldsymbol{k}) \exp\big[2\pi j\boldsymbol{k}^T\boldsymbol{N}^{-1}\boldsymbol{n}\big] \tag{5a}$$

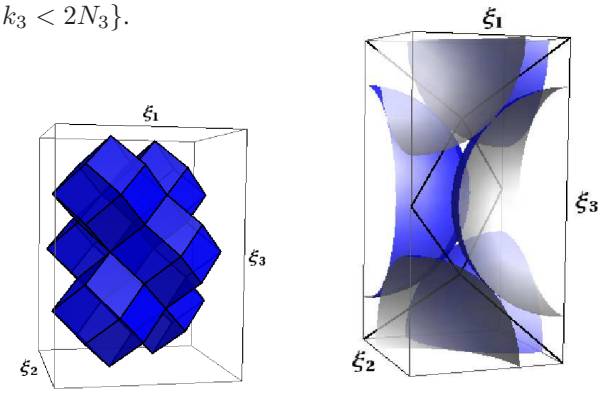
$$f_1(\boldsymbol{n}) = \frac{1}{N}\sum_{\boldsymbol{k}\in\mathcal{K}} F(\boldsymbol{k}) \exp\big[2\pi j\boldsymbol{k}^T\boldsymbol{N}^{-1}(\boldsymbol{n} + \frac{1}{2}\boldsymbol{t})\big] \tag{5b}$$

where $N = 2N_1N_2N_3$ is the number of samples and $\mathcal{K} \subset \mathbb{Z}^3$ is any set that indexes all the distinct frequency samples. It is easily verified that both the sequences (5a) and (5b) are periodic with periodicity matrix $\boldsymbol{N}$.

3.1.1 Efficient Evaluation

Since $\boldsymbol{N}$ is diagonal, the kernel in both equations (4) and (5) is separable. This suggests that the transform can be efficiently computed via the rectangular multidimensional FFT, provided that a suitable rectangular index set $\mathcal{K}$ can be found. Observe that the Cartesian sequence $F(\boldsymbol{k})$ is periodic with periodicity matrix $2\boldsymbol{N}$, i.e $F(\boldsymbol{k} + 2\boldsymbol{N}\boldsymbol{r}) = F(\boldsymbol{k})$ for all $\boldsymbol{r} \in \mathbb{Z}^3$. Therefore, one way to obtain a rectangular index set is to choose $\mathcal{K}$ such that it contains all the frequency indices within one period generated by the matrix $2\boldsymbol{N}$. This consists of a total of $|\det 2\boldsymbol{N}| = 4N$ indices and hence contains four replicas of the fundamental rhombic dodecahedron.

A non-redundant rectangular index set can be found by exploiting the geometric properties of the FCC lattice. If we consider the first octant only, $4N$ samples are contained within a cube formed by the FCC lattice sites that have even parity. This cube also contains six face-centered sites. By joining any two axially opposite face-centered sites, we can split the cube into four rectangular regions such that each region consists of non-redundant samples only. Six rhombic dodecahedra contribute to such a region as illustrated in Fig. 2. The non-redundant region shown in Fig. 2b is obtained by limiting $\boldsymbol{k}$ to the index set given by $\mathcal{K} = \{\boldsymbol{k} \in \mathbb{Z}^3 : 0 \le k_1 < N_1, 0 \le k_2 < N_2, 0 \le k_3 < 2N_3\}$.



(a) (b)

Figure 2: (a) Six rhombic dodecahedra contribute to a non-redundant rectangular region. (b) Zoomed in view of the non-redundant rectangular region that contains the full spectrum split into six pieces. ξ_1, ξ_2 and ξ_3 indicate the principal directions in the frequency domain.

This region can further be subdivided into two cubes stacked on top of each other, each containing $N_1 \times N_2 \times N_3$ samples. The forward transform (4) can then be evaluated in the two cubes separately by appropriately applying the Cartesian FFT to the two sequences $f_0(\boldsymbol{n})$ and $f_1(\boldsymbol{n})$ and combining the results together. After rearranging terms in (4), the forward transform in the bottom cube becomes

$$F_0(\boldsymbol{k}) = F(\boldsymbol{k}) = \sum_{\boldsymbol{n}\in\mathcal{N}} f_0(\boldsymbol{n}) \exp\big[-2\pi j \boldsymbol{k}^T \boldsymbol{N}^{-1}\boldsymbol{n}\big] + \exp[-\pi j \boldsymbol{k}^T \boldsymbol{N}^{-1}\boldsymbol{t}] \sum_{\boldsymbol{n}\in\mathcal{N}} f_1(\boldsymbol{n}) \exp\big[-2\pi j \boldsymbol{k}^T \boldsymbol{N}^{-1}\boldsymbol{n}\big], \quad (6)$$

where $\boldsymbol{k}$ is now restricted to the set $\mathcal{N}$. Since this equation is valid for all $\boldsymbol{k} \in \mathbb{Z}^3$, the forward transform in the top cube can be computed from (6) by $F_1(\boldsymbol{k}) = F_0(\boldsymbol{k} + (0,0,N_3)^T)$ which simplifies to

$$F_1(\boldsymbol{k}) = \sum_{\boldsymbol{n}\in\mathcal{N}} f_0(\boldsymbol{n}) \exp\big[-2\pi j \boldsymbol{k}^T \boldsymbol{N}^{-1}\boldsymbol{n}\big] - \exp[-\pi j \boldsymbol{k}^T \boldsymbol{N}^{-1}\boldsymbol{t}] \sum_{\boldsymbol{n}\in\mathcal{N}} f_1(\boldsymbol{n}) \exp\big[-2\pi j \boldsymbol{k}^T \boldsymbol{N}^{-1}\boldsymbol{n}\big], \quad (7)$$

for $\boldsymbol{k} \in \mathcal{N}$. Equations (6) and (7) are now in a form that permits a straightforward application of the Cartesian FFT. Since the two equations are structurally similar, only two $N_1 \times N_2 \times N_3$ FFT computations are needed, one for the sequence $f_1(\boldsymbol{n})$ and one for $f_2(\boldsymbol{n})$.

In a similar fashion, the inverse transform (5) can be computed using two inverse FFT computations. Splitting the summations in (5) into the two constituent cubes gives

$$f_0(\boldsymbol{n}) = \frac{1}{N} \sum_{\boldsymbol{k}\in\mathcal{N}} (F_0(\boldsymbol{k}) + F_1(\boldsymbol{k})) \exp\big[2\pi j \boldsymbol{k}^T \boldsymbol{N}^{-1}\boldsymbol{n}\big],$$

$$f_1(\boldsymbol{n}) = \frac{1}{N} \sum_{\boldsymbol{k}\in\mathcal{N}} \Big((F_0(\boldsymbol{k}) - F_1(\boldsymbol{k})) \exp[\pi j \boldsymbol{k}^T \boldsymbol{N}^{-1}\boldsymbol{t}]\Big) \cdot \exp\big[2\pi j \boldsymbol{k}^T \boldsymbol{N}^{-1}\boldsymbol{n}\big]. \quad (8)$$

3.2 FCC DFT

The FCC lattice with arbitrary scaling is generated by the sampling matrix $h\boldsymbol{L}_{\mathsf{F}}$. The rhombic dodecahedral Voronoi cell has a volume of $|\det h\boldsymbol{L}_{\mathsf{F}}| = 2h^3$. The frequency spectrum is replicated according to (3) on a reciprocal BCC lattice that has a truncated octahedral Voronoi cell having a volume of $\frac{1}{2h^3}$.

A sequence sampled on the FCC lattice can be split up into four Cartesian subsequences corresponding to the four Cartesian cosets. Each subsequence is given by

$$f_i(\boldsymbol{n}) = f_c(2h\boldsymbol{I}\boldsymbol{n} + h\boldsymbol{t}_i),$$

where $i \in \{0,1,2,3\}$ and $\boldsymbol{t}_i$ are the integer shift vectors $(0,0,0)^T$, $(1,0,1)^T$, $(0,1,1)^T$ and $(1,1,0)^T$ respectively. Analogous to the BCC case, let us choose a rectangular truncation of the original sequence by limiting $\boldsymbol{n}$ to the set $\mathcal{N}$ and extend the sequences periodically so that they satisfy $f_i(\boldsymbol{n}+\boldsymbol{N}\boldsymbol{r}) = f_i(\boldsymbol{n})$. This truncation yields a rectangular fundamental region in the spatial domain consisting of a total of $N = 4N_1N_2N_3$ distinct samples. Therefore, each truncated octahedron in the frequency domain tessellation will consist of N distinct points that are sampled in a Cartesian fashion at the frequencies $\boldsymbol{\xi} = \frac{1}{2h}\boldsymbol{N}^{-1}\boldsymbol{k}$ where $\boldsymbol{k} \in \mathbb{Z}^3$. The sampled sequence in the frequency domain is thus given by

$$F(\boldsymbol{k}) = \hat{F}(\frac{1}{2h}\boldsymbol{N}^{-1}\boldsymbol{k}) = \sum_{\boldsymbol{n}\in\mathcal{N}} \sum_{i=0}^{3} f_i(\boldsymbol{n}) \exp\big[-2\pi j \boldsymbol{k}^T \boldsymbol{N}^{-1}(\boldsymbol{n} + \frac{1}{2}\boldsymbol{t}_i)\big]. \quad (9)$$

This defines a forward FCC DFT. Like the BCC case, it is invariant under shifts of the type $\boldsymbol{\xi} = \frac{1}{2h}(\boldsymbol{N}^{-1}\boldsymbol{k} + \boldsymbol{L}_{\mathsf{B}}\boldsymbol{r})$ making it periodic on a BCC lattice with one fundamental period contained in a truncated octahedron.

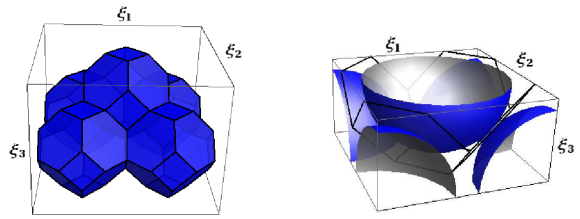
The inverse FCC DFT is obtained by summing over all the distinct sinusoids evaluated at the spatial sample locations

$$f_i(\boldsymbol{n}) = \frac{1}{N} \sum_{\boldsymbol{k}\in\mathcal{K}} F(\boldsymbol{k}) \exp\big[2\pi j \boldsymbol{k}^T \boldsymbol{N}^{-1}(\boldsymbol{n} + \tfrac{1}{2}\boldsymbol{t}_i)\big], \quad (10)$$

where $\mathcal{K} \subset \mathbb{Z}^3$ is any set that indexes all the N distinct sinusoids.

3.2.1 Efficient Evaluation

Since $\boldsymbol{N}$ is diagonal, the key to efficiently evaluating the FCC DFT pair (9) and (10) is to choose a suitable rectangular region in the frequency domain that contains N distinct samples. Similar to the BCC DFT, the sequence (9) is $2\boldsymbol{N}$ periodic with one complete rectangular period containing $|\det 2\boldsymbol{N}| = 2N$ samples and hence two complete spectrum replicas. These $2N$ samples are contained within a cube, the corners of which lie at the even parity points of the BCC lattice. This cubic region can be split into two by halving along any of the three principal directions yielding a rectangular region that contains only non-redundant samples as illustrated in Fig. 3. The index set that spans the region depicted in Fig. 3b is given by $\mathcal{K} = \{\boldsymbol{k} \in \mathbb{Z}^3 : 0 \leq k_1 < 2N_1, 0 \leq k_2 < 2N_2, 0 \leq k_3 < N_3\}$.



(a) (b)

Figure 3: (a) Five truncated octahedra contribute to a non-redundant rectangular region. (b) Zoomed in view of the rectangular region that contains the full spectrum.

The non-redundant region can be split into four $N_1 \times N_2 \times N_3$ cubic subregions and the forward transform (9) can be evaluated in each of the subregions separately by appropriately applying the FFT to each of the subsequences $f_i(\boldsymbol{n})$ and combining the output. The derivation is very similar to the BCC case and we leave the details to the reader. The forward transform in each subregion can be written as

$$F_m(\boldsymbol{k}) = \sum_{i=0}^{3} H_{im} \exp[-\pi j \boldsymbol{k}^T \boldsymbol{N}^{-1} \boldsymbol{t}_i] \cdot \Big(\sum_{\boldsymbol{n} \in \mathcal{N}} f_i(\boldsymbol{n}) \exp\big[-2\pi j \boldsymbol{k}^T \boldsymbol{N}^{-1} \boldsymbol{n}\big] \Big), \tag{11}$$

where $m \in \{0, 1, 2, 3\}$, $\boldsymbol{k} \in \mathcal{N}$ and H_{im} is an element of the 4×4 Hadamard matrix $\boldsymbol{H} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$. The four subregions $F_m(\boldsymbol{k})$ have their bottom left corners at the frequency index vectors $(0,0,0)^T$, $(N_1,0,0)^T$, $(0,N_2,0)^T$ and $(N_1,N_2,0)$ respectively.

Likewise, the inverse transform (10) can be evaluated using four inverse FFT evaluations, one for each of the subsequences. This yields

$$f_i(\boldsymbol{n}) = \frac{1}{N} \sum_{\boldsymbol{k} \in \mathcal{N}} \big(\exp[\pi j \boldsymbol{k}^T \boldsymbol{N}^{-1} \boldsymbol{t}_i] \sum_{m=0}^{3} H_{im} F_m(\boldsymbol{k}) \big) \cdot \exp\big[2\pi j \boldsymbol{k}^T \boldsymbol{N}^{-1} \boldsymbol{n}\big]. \tag{12}$$

4. Discussion

The decomposition of the non-redundant region in the frequency domain into cubes leads to transforms that are much more efficient. Both the BCC and FCC DFTs proposed by Csébfalvi et al. [1] are redundant and require the FFT of a $2N_1 \times 2N_2 \times 2N_3$ sequence. In contrast, our proposed evaluation strategy eliminates the redundancy and computes only two $N_1 \times N_2 \times N_3$ FFTs for the BCC case and four $N_1 \times N_2 \times N_3$ FFTs for the FCC case.

Any operation in the frequency domain must respect the arrangement of the different portions of the spectrum. The BCC DFT splits the spectrum into six parts as illustrated by the six pieces (two lunes and four spherical triangles) of the sphere in Fig. 2b. The FCC transform splits the frequency spectrum into five parts as indicated by the hemisphere and the four spherical triangles in Fig. 3b.

5. Summary

In this paper, we have shown that a MDFT of a Cartesian periodic sequence sampled on the BCC or FCC lattices can be efficiently evaluated using the FFT. The BCC lattice can be represented as two shifted Cartesian lattices. This representation leads to a separable transform that is efficiently computed via two non-redundant FFT evaluations of the Cartesian subsequences. Similarly, the FCC lattice consists of four shifted Cartesian lattices and the MDFT requires four non-redundant FFT evaluations.

References:

[1] B. Csébfalvi and B. Domonkos. Pass-Band Optimal Reconstruction on the Body-Centered Cubic Lattice. In *Vision, Modeling, and Visualization 2008: Proceedings, October 8-10, 2008, Konstanz, Germany*, page 71. IOS Press, 2008.

[2] A. Entezari. *Optimal Sampling Lattices and Trivariate Box Splines*. PhD thesis, Simon Fraser University, Vancouver, Canada, July 2007.

[3] A. Entezari, D. Van De Ville, and T. Möller. Practical box splines for volume rendering on the body centered cubic lattice. *IEEE Transactions on Visualization and Computer Graphics*, 14(2):313 – 328, 2008.

[4] A. Guessoum and R. Mersereau. Fast algorithms for the multidimensional discrete Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-34(4):937–943, 1986.

[5] M. Kim, A. Entezari, and J. Peters. Box Spline Reconstruction on the Face Centered Cubic Lattice. *IEEE Transactions on Visualization and Computer Graphics (Proceedings Visualization/Information Visualization 2008)*, 14(6):1523–1530, 2008.

[6] R. Mersereau. The Processing of Hexagonally Sampled Two-dimensional Signals. *Proceedings of the IEEE*, 67(6):930–949, June 1979.

[7] R. Mersereau and T. Speake. The processing of periodically sampled multidimensional signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, (1):188–194, 1983.

[8] T. Theußl, T. Möller, and M. Gröller. Optimal regular volume sampling. In *Proceedings of the conference on Visualization'01*, pages 91–98. IEEE Computer Society Washington, DC, USA, 2001.

[9] P. Vaidyanathan. Fundamentals of multidimensional multirate digital signal processing. *Sadhana*, 15(3):157–176, 1990.

Daubechies Localization Operator in Bargmann - Fock Space and Generating Function of Eigenvalues of Localization Operator

Kunio Yoshino, Tamazutsumi, 1-28-1, Setagaya-ku, Tokyo, Japan,158-8557.
yoshinok@tcu.ac.jp.

Abstract:

We will express Daubechies localization operators in Bargmann - Fock space. We will prove that the Hermite functions are eigenfunctions of Daubechies localization operator. By making use of generating function of eigenvalues of Daubechies localization operator, we will show some reconstruction formulas for symbol function of Daubechies localization operator with rotational invariant symbol.

1. Introduction

Daubechies localization operator was introduced in
I. Daubechies : A Time Frequency Localization Operator: A Geometric Phase Space Approach, IEEE. Trans. Inform. theory. vol.34, pp.605-612(1988)

She obtained following results.

Theorem(Daubechies)([2])
Suppose that symbol function of Daubechies localization operator is rotational invariant. Then
(i) Eigenfunctions of Daubechies localization operator are Hermite functions.
(ii) Eigenvalues are given by Mellin transform of symbol function.

In this paper we realize Daubechies localization opeartor in Bargamann - Fock space. We will consider the eigenvalue problem of Daubechies localization opeartor in Bargmann - Fock space. By making use of Bargamann - Fock space we will give a new proof of above theorem. We will establish reconstruction formula of symbol function of Daubechies localization operator with rotational invariant symbol by generating function of eigenvalues of Daubechies localization operator. For the simplicity, we will confine ourselves to 1-dimensional case.

2. Bargmann Transform

Put
$$A(z,x) = \pi^{-1/4} \exp\left\{-\frac{1}{2}(z^2+x^2) + \sqrt{2}z\cdot x\right\},$$
where $z \in \mathbb{C}$ and $x \in \mathbb{R}$.

Bargmann transform $B(\psi)$ is defined as follows :

$$B(\psi)(z) \stackrel{def}{=} \int_{\mathbb{R}} \psi(x)A(z,x)dx, \quad (\psi \in L^2(\mathbb{R})).$$

Put

$$BF = \{g \in H(\mathbb{C}) : \int_{\mathbb{C}} |g(z)|^2 e^{-|z|^2} dz \wedge d\bar{z} < \infty\}$$

where $H(\mathbb{C})$ denotes the space of entire functions in the complex plane.

BF is called Bargmann-Fock space.
Theorem 1([1])

Bargmann transform is a unitary mapping from $L^2(\mathbb{R})$ to Bargmann-Fock space BF.

For the details of Bargmann transform and Bargmann - Fock space, we will refer the reader to [1] and [3].

3. Hermite Functions

Definition 1([1],[3]) Hermite functions $h_m(x)$ is defined by :

$$h_m(x) = (-1)^m(2^m m!\sqrt{\pi})^{-1/2}\exp(x^2/2)\frac{d^m}{dx^m}\exp(-x^2),$$
$(m \in \mathbb{N})$.
Hermite functions has following generating function expansion :

$$\pi^{-1/4}\exp\left\{-\frac{1}{2}(z^2+x^2)+\sqrt{2}z\cdot x\right\}$$
$$= \sum_{m=0}^{\infty}\frac{z^m}{\sqrt{m!}}h_m(x),$$
$(z \in \mathbb{C}, x \in \mathbb{R})$.

We recall some well known facts about Hermite functions.
Proposition 1([1],[3])

(i) $\{h_m(x)\}_{m=0}^{\infty}$ is complete orthonormal basis in $L^2(\mathbb{R})$.

(ii) $(-\frac{\partial^2}{\partial x^2}+x^2-1)h_m(x)=mh_m(x),$

(iii) $B(h_m)(z)=\frac{z^m}{\sqrt{m!}}, \quad (z\in\mathbb{C})$

(iv) $\mathfrak{F}(h_m)(x)=(-i)^m h_m(x),$

where $\mathfrak{F}$ is Fourier transform.

Proposition 2([1],[3])

(i) $(B\circ L\circ B^{-1})g(z)=z\frac{\partial}{\partial z}g(z),$
where $L=-\frac{\partial^2}{\partial x^2}+x^2-1.$

(ii) $(B\circ\mathfrak{F}\circ B^{-1})g(z)=g(-iz),$

where $\mathfrak{F}$ is Fourier transform and $g(z)\in BF$.

4. Daubechies Localization Operator

Put
$\phi_{p,q}(x)=\pi^{-1/4}e^{ipx}e^{-(x-q)^2/2}.$

$$<\phi_{p,q},f>=\int_{\mathbb{R}}\phi_{p,q}(x)f(x)dx.$$

This is so called Short time Fourier transform (or Windowed Fourier transform, or Gabor transform).
Definition 2([2])
Suppose that $F(p,q)\in L^1(\mathbb{R}^2)$ and $f(x)\in L^2(\mathbb{R})$.
We put

$$P_F(f)(x)=\frac{1}{2\pi}\int\int_{\mathbb{R}^2}F(p,q)\phi_{p,q}(x)<\phi_{p,q},f>dpdq,$$

We call P_F (Daubechies) localization operator $F(p,q)$ is called symbol function.
Daubechies obtained following results.
Theorem([2]). Suppose that $F(p,q)\in L^1(\mathbb{R}^2)$ and

$F(p,q)$ is rotational invariant function,
i.e. $F(p,q)=\tilde{F}(r^2), \quad (r^2=p^2+q^2).$
Then
(i) Hermite functions $h_m(x)$ are eigenfunctions of Daubechies operator P_F.

$$P_F(h_m)(x)=\lambda_m h_m(x), \quad (m\in\mathbb{N}),$$

(ii) $\lambda_m=\frac{1}{m!}\int_0^{\infty}e^{-s}s^m\tilde{F}(2s)ds, \quad (m\in\mathbb{N}).$

5. A Realization of Daubechies Localization Operator in Bargmann Fock space

In this section we will express Daubechies Localization Operator in Bargmann - Fock space.
First we need following lemmas.
Lemma 1

$$B(\phi_{p,q})(z)=e^{zw-1/2|w|^2+1/2ipq}, \quad (w=\frac{p+iq}{\sqrt{2}})$$

Lemma 2([1])

$$g(z)=\frac{1}{2\pi i}\int\int_{\mathbb{C}}e^{w\bar{t}}g(t)e^{-|t|^2}dt\wedge d\bar{t}, \quad (g\in BF)$$

Theorem 2 Under the same assumptions in Prop. 3, we have
$(B\circ P_F\circ B^{-1})(g)(z)$
$=\frac{1}{2\pi i}\int\int_{\mathbb{C}}F(w,\overline{w})e^{z\overline{w}}g(w)e^{-|w|^2}dw\wedge d\overline{w},$
$(\forall g\in BF)$

(**Proof**)
Since Bargmann transform is unitary operator, we have
$P_F(f)(x)=\frac{1}{2\pi}\int\int F(p,q)\phi_{p,q}(x)<\phi_{p,q},f>dpdq,$
$=\frac{1}{2\pi}\int\int F(p,q)\phi_{p,q}(x)<B\phi_{p,q},Bf>dpdq,$
So by lemma 1,
$B\circ P_F(f)(x)$
$=\frac{1}{2\pi}\int\int F(p,q)B\phi_{p,q}(z)<B\phi_{p,q},Bf>dpdq,$
$=\frac{1}{2\pi}\int\int F(p,q)e^{zw-1/2|w|^2+1/2ipq}<B\phi_{p,q},Bf>dpdq,$
Hence we have
$(B\circ P_F\circ B^{-1})(g)(z)$
$=\frac{1}{2\pi}\int\int F(p,q)e^{zw-1/2|w|^2+1/2ipq}<B\phi_{p,q},g>dpdq,$
On the other hand
$<B\phi_{p,q},g>$
$=\frac{1}{2\pi}\int\int e^{t\bar{w}-1/2|w|^2-1/2ipq}g(t)e^{-|t|^2}dtd\bar{t},$
By Lemma 2,
$=e^{-1/2|w|^2-1/2ipq}g(\bar{w})$
Thus we obtained our desired result.
Proposition 3([2]). Suppose that $F(p,q)\in L^1(\mathbb{R}^2)$ and

$F(p,q)$ is rotational invariant function,
i.e. $F(p,q)=\tilde{F}(r^2), \quad (r^2=p^2+q^2).$
Then
(i) Functions z^m are eigenfunctions of operator $B\circ P_F\circ B^{-1}$.

$$(B\circ P_F\circ B^{-1})(z^m)=\lambda_m z^m, \quad (m\in\mathbb{N}),$$

(ii) $\lambda_n=\frac{1}{n!}\int_0^{\infty}e^{-s}s^n\tilde{F}(2s)ds, \quad (n\in\mathbb{N}).$

(Proof)

By Theorem 2, we have
$(B \circ P_F \circ B^{-1})(z^m)$
$= \frac{1}{2\pi i} \int \int_{\mathbb{C}} F(2|w|^2) e^{z\overline{w}} w^m e^{-|w|^2} dw \wedge d\overline{w},$

Employing polar coordinte transform $w = re^{i\theta}$ and $s = r^2$,
we have
$= z^m \frac{1}{m!} \int_0^\infty e^{-s} s^m \tilde{F}(2s) ds.$

As a corollary of Proposition 3, we obtained following Daubechies's results in section 4.

Proposition 4([8]) Let $\{\lambda_m\}$ be eigenvalues of P_F. Then
there exists a positive constant C such that
$|\lambda_m| \leq \frac{C}{\sqrt{|m|}}, \quad (m \in \mathbb{N}).$

Put
$\Lambda(w) = \sum_{m=0}^{\infty} \lambda_m w^m.$
We call $\Lambda(w)$ generating function of eigenvalues of Daubechies Localization Operator.
Theorem 3 Under the same assumptions in Prop. 3, we have

$$(B \circ P_F \circ B^{-1})(g)(z) = (2\pi i)^{-n} \oint g(t) \Lambda(\frac{z}{t}) \frac{dt}{t},$$

$(\forall g \in BF)$
(Proof)
Suppose that $g(z) \in BF$. We consider Taylor expansion of $g(z)$ at the origin.
Put
$g(z) = \sum_{m=0}^{\infty} a_m z^m$
By Proposition 3, we have
$(B \circ P_F \circ B^{-1})(z^m) = \lambda_m z^m.$
So
$(B \circ P_F \circ B^{-1})(g)(z) = (B \circ P_F \circ B^{-1})(\sum_{m=0}^{\infty} a_m z^m)$
$= \sum_{m=0}^{\infty} a_m \lambda_m z^m = (2\pi i)^{-n} \oint g(t) \Lambda(\frac{z}{t}) \frac{dt}{t}$
Hence we have
$(B \circ P_F \circ B^{-1})(g)(z) = (2\pi i)^{-1} \oint g(t) \Lambda(\frac{z}{t}) \frac{dt}{t}.$

6. An Example of Daubechies Localization Operator

In this section we will consider following special Daubechies localization operators.

Put
$F_a(p,q) = e^{\frac{a-1}{2a}(p^2+q^2)} = e^{\frac{a-1}{2a} r^2}, \quad (0 < a < 1).$
Then
$\lambda_m = a^{m+1}, \quad \Lambda(w) = \frac{a}{1-aw}.$
$P_{F_a}(h_m)(x) = a^{m+1} h_m(x).$
$P_{F_a} = \sum_{m=0}^{\infty} a^{m+1} h_m(x) h_m(y).$
valids in operator sense.
$(P_{F_a} = \sum_{m=0}^{\infty} a^{m+1} |m><m|, \quad$ in Dirac's Notation.)

If $a = 2^{-1}$, this is Schatten decomposition of P_{F_a} and P_{F_a} is called density operator in quantum statistical mechanics.

Proposition 5 (Mehler's formula [3],[5])

$\sum_{m=0}^{\infty} a^{m+1} h_m(x) h_m(y)$
$= \frac{a}{\sqrt{\pi(1-a^2)}} e^{\frac{-1}{4}(\frac{1-a}{1+a}(x+y)^2 + \frac{1+a}{1-a}(x-y)^2)}, \quad (|a| < 1).$

Corollary 3
(i) $P_{F_a}(f)$
$= \int_{\mathbb{R}} \frac{a}{\sqrt{\pi(1-a^2)}} e^{\frac{-1}{4}(\frac{1-a}{1+a}(x+y)^2 + \frac{1+a}{1-a}(x-y)^2)} f(y) dy,$
$(f \in L^2).$

(ii) If $a \in \mathbb{C}, |a| < 1$, then
$P_{F_a} : L^2 \longrightarrow L^2$ is bounded linear operator.
(Proof)
If $a \in \{a \in \mathbb{C} : |a| < 1\}$, then real part of $\frac{1-a}{1+a} + \frac{1+a}{1-a}$ is positive. So P_{F_a} is bounded linear operator from L^2 to L^2. Namely, we obtained analytic continuation of P_{F_a} under the condition $(a \in \mathbb{C}, |a| < 1)$.

7. Realization of P_{F_a} in Bargmann - Fock space

In this section we will consider P_{F_a} in Bargmann - Fock space.

Proposition 6
(i) $B \circ P_{F_a} \circ B^{-1} = \sum_{m=0}^{\infty} a^{m+1} \frac{z^m}{\sqrt{m!}} \frac{\bar{w}^m}{\sqrt{m!}}.$
valids in operator sense.

(ii) $(B \circ P_{F_a} \circ B^{-1})(g)(z)$
$= \frac{ia}{2} \int \int_{\mathbb{C}} e^{az\bar{w}} g(w) e^{-|w|^2} dw \wedge d\bar{w},$
$(g \in BF)$
(Proof)
Since $\frac{z^m}{\sqrt{m!}}$. are eigenfunctions of $B \circ P_{F_a} \circ B^{-1}$,
we have
$B \circ P_{F_a} \circ B^{-1} = \sum_{m=0}^{\infty} a^{m+1} \frac{z^m}{\sqrt{m!}} \frac{\bar{w}^m}{\sqrt{m!}}.$

Proposition 7 Suppose that $|a| < 1, (a \in \mathbb{C})$. Then we have

$$(B \circ P_{F_a} \circ B^{-1})(g)(z) = a\, g(az), \quad (g \in BF).$$

(Proof)
$(B \circ P_{F_a} \circ B^{-1})(g)(z) = (2\pi i)^{-1} \oint g(t)\Lambda(\frac{z}{t})\frac{dt}{t}$
$= (2\pi i)^{-1} \oint g(t)\frac{a}{t-az}dt = ag(az).$

Proposition 8 For $f \in L^2$, we have

(i) $\lim_{a\to 1} P_{F_a}(f) = f,$
(ii) $\lim_{a\to -i} P_{F_a}(f) = (-i)\mathfrak{F}f,$

(iii) $\lim_{a\to i} P_{F_a}(f) = i\mathfrak{F}^{-1}f,$

where $\mathfrak{F}$ is Fourier transform.

(Proof) By Prop.7, we have

$$(B \circ P_{F_a} \circ B^{-1})(g)(z) = ag\,(az)\,, \quad (g \in BF).$$

(i) $\lim_{a\to 0}(B \circ P_{F_a} \circ B^{-1})(g) = \lim_{a\to 1} ag(az) = g(z).$
This means that $\lim_{a\to 1} P_{F_a} =$ Identity operator.
(ii) $\lim_{a\to -i}(B \circ P_{F_a} \circ B^{-1})(g) = \lim_{a\to -i} ag(az)$
$= (-i)g(-iz).$
By (ii) in Proposition 2, this means that
$\lim_{a\to -i} P_{F_a} = (-i)\mathfrak{F}$.
Proof of (iii) is same as that of (ii).

Proposition 9
(i)
$$G = \{PF_a : a \in \mathbb{C}, |a| < 1\} \cup \{I_d\}$$
is semigroup.
(ii)
$$P_{F_a} \circ P_{F_a} = P_{F_{ab}}.$$

(Proof) By Proposition 7,
$(B \circ P_{F_a} \circ B^{-1})(g)(z) = ag(az), \quad g(z) \in BF$
So, we have
$(B \circ P_{F_a} \circ P_{F_b} \circ B^{-1})(g)(z) = bag\,(baz)$
Hence we have
$P_{F_b} \circ P_{F_a} = P_{F_{ab}}.$

In these cases, $F_a(p,q) \notin L^1$. But these operators still define bounded operators from L^2 to L^2.
As seen in Proposition 8, these operators are obtained as limit of $PF_a, (F_a \in L^1)$.

8. *Reconstruction formulas*

We assume that $F(p,q)$ is rotational invariant L^1 function. Namely, $F(p,q) = \tilde{F}(\sqrt{p^2+q^2})$.

In section 5, we introduced following generating function:
$\Lambda(w) = \sum_{m=0}^{\infty} \lambda_m w^m$
$\Lambda(w)$ is called generating function for eigenvalues of P_F
Now we consider following formal power series :

$$\sum_{m=0}^{\infty} m!\lambda_m t^{-m-1}$$

In general this series is divergent series. We put
$G(t) = \int_0^{\infty} \frac{\tilde{F}(2s)e^{-s}}{t-s}ds, \quad (t \in \mathbb{C}\backslash[0,\infty]).$
We have
Proposition 10([8])
Formal power series
$\sum_{m=0}^{\infty} m!\lambda_m t^{-m-1}$
is an asymptotic expansion of $G(t)$.

Remark $\Lambda(w)$ is the Borel transform of formal power series $\sum_{m=0}^{\infty} m!\lambda_m t^{-m-1}$.

Since $G(t)$ is Hilbert transform of $\tilde{F}(2s)e^{-s}$, we have
Theorem 5
$\tilde{F}(2s) = e^s \lim_{t\to 0} \frac{-1}{2\pi i}\left(G(s+it) - G(s-it)\right)$
We also have
Theorem 6([8])

$$\tilde{F}(2s) = (2\pi)^{-n} e^s \mathfrak{F}(\Lambda(iv))(s),$$

valids in distribution sense.
where $\mathfrak{F}$ is Fourier transform.

References:

[1] V. Bargmann : *On a Hilbert Space of Analytic Functions and an Associated Integral Transform Part I,* Comm.Pure.Appl.Math, pp. 187-214(1961)

[2] I. Daubechies : *A time frequency localization operator; A geometric phase space approach,* IEEE. Trans. Inform. theory. vol.34, pp.605-612(1988)

[3] G. B. Folland : *Harmonic Analysis in Phase Space,* Princeton Univ. Press (1989)

[4] K. Gröhenig: *Foundations of Time-Frequency Analysis*, Birkhäuser-Verlag, Basel, Berlin, Boston(2000)

[5] M.W. Wong : *Weyl Transforms,* Springer-Verlag. New York. (1998)

[6] M.W. Wong : *Localization Operators on the Weyl-Heisenberg Group,* Geometry, Analysis and Applications, Proceedings of the International Conference (editor:P.S.Pathak) 303-314(2001)

[7] M.W. Wong : *Wavelet Transforms and Localization Operator,* Birkhäuser-Verlag. Basel, Berlin, Boston. (2002)

[8] K. Yoshino : *Localization operators in Bargmann - Fock space and reconstruction formula for symbol functions*, preprint (2009)

Signal-dependent sampling and reconstruction method of signals with time-varying bandwidth

Modris Greitans and Rolands Shavelis

Institute of Electronics and Computer Science, 14 Dzerbenes str., Riga LV-1006, Latvia.
greitans@edi.lv, shavelis@edi.lv

Abstract:

The paper describes the sampling method of nonstationary signals with time-varying spectral bandwidth. The reconstruction procedure exploiting the low-pass filter with time-varying cut-off frequency is derived. The filter application in signal reconstruction from its level-crossing samples is shown. The results of computer simulations are presented.

1. Introduction

The spectral characteristics of signals of practical interest often change with time. Generally, a signal with time-varying spectral bandwidth can be approximated with fewer samples per interval using appropriate non-equidistantly spaced samples than using uniform sampling procedure, where the sampling rate is chosen taking into account the highest signal frequency. For example, let us inspect a signal with wide bandwidth regions and narrow spectral bandwidth in the rest of signal observation. It is more efficient to sample the narrow bandwidth regions at a lower rate than the regions, where spectral bandwidth is wide. Solving this problem correctly requires the knowledge of the function of the instantaneous maximum frequency of signal. The paper will show two typical situations. First, information about the time-varying bandwidth is known a priori. In this case the deliberately non-uniform sampling instants can be calculated in advance, and reconstruction is based on application of filter with appropriate time-varying impulse response function. Second, the signal-dependent sampling scheme - level crossing sampling (LCS) is used for analog-to-digital (A/D) conversion. The idea of level-crossing sampling is based on the principle that samples are captured when the input signal crosses predefined levels. Such a sampling strategy has quite long history and is exploited for various applications [1, 2]. It has been shown that LCS has several interesting properties and is more efficient than traditional sampling in many respects [3]. In particular, it can be related to the processing of non-stationary signals, because if a waveform is changing rapidly, the samples are spaced more closely, and conversely – if a signal is varying slowly, the samples are spaced sparsely [4]. This property allows to calculate the estimate of the function of the instantaneous maximum frequency of signal from the positions of samples. In this case to reconstruct the waveform of signal, an additional resampling procedure is needed before the use of time-varying reconstruction filter, which will be described in next section.

Note that in both cases the local sampling density reflects the local bandwidth of the signal, therefore samples are spaced non-uniformly and advanced algorithms are required for digital signal processing.

2. Reconstruction of signal with time-varying bandwidth

There are several methods used for reconstruction of non-uniformly sampled band-limited signals. For correct recovery, they typically require that the maximal length of the gaps between the sampling instants does not exceed the Nyquist rate [5]. If the signal is non-stationary with time-varying spectral bandwidth, satisfying globally this requirement is not an appropriate decision, because this provides redundant data. The use of level-crossing sampling scheme can reduce the amount of samples, because the intervals between samples are determined by signal local properties and by the number of quantization levels. The quality of processing can be improved if the recovery procedure takes into account the local bandwidth of the signal [6]. In the following subsections the proposed idea and methods for reconstruction using filters with time-varying bandwidth and for the estimation of local maximum frequency of signal from its level-crossing samples will be discussed.

2.1 Idea of signal-dependent reconstruction functions

The sampling theorem states that every bandlimited signal $s(t)$ can be reconstructed from its equidistantly spaced samples if the sampling rate equals or exceeds the Nyquist rate $2F_{max}$, where F_{max} is the maximum frequency in the signal spectrum. The reconstruction in time domain can be expressed as

$$\hat{s}(t) = \sum_{n=0}^{N-1} s(t_n) h(t - t_n), \qquad (1)$$

where $\widehat{s}(t)$ denotes reconstructed signal, N is the number of the original signal samples $s(t_n)$ and $h(t)$ is an appropriate impulse response of the reconstruction filter, classi-

cally, sinc-function

$$h_1(t) = \text{sinc}(2\pi F_{max} t) \tag{2}$$

As the sampling instants $t_n = \frac{n}{2F_{max}}$, then the impulse response

$$h_1(t - t_n) = h_1(t, t_n) = \text{sinc}(2\pi F_{max} t - n\pi), \tag{3}$$

where $h_1(t - t_n) = h(t, t_n)$ is written as the function of two arguments. The reconstructed signal becomes

$$\hat{s}(t) = \sum_{n=0}^{N-1} s(t_n) h_1(t, t_n) \tag{4}$$

If the signal with time-varying frequency bandwidth $f_{max}(t)$ is considered, then the sampling rate of the signal according to Nyquist must be at least $2F_{max}$, where $F_{max} = \max(f_{max}(t))$. In this case any information about the local spectral bandwidth is ignored during the sampling process. To take it into account, it is proposed instead of $h_1(t, t_n)$ to use more general function

$$h_2(t, t_n) = \text{sinc}(\Phi(t) - \Phi(t_n)) = \text{sinc}(\Phi(t) - n\pi), \tag{5}$$

where $\Phi(t) = 2\pi \int_0^t f_{max}(t)dt$ is the phase of the sinusoid, whose frequency changes in time as $f_{max}(t)$, $t \geq 0$ and sampling instants t_n are chosen such that $\Phi(t_n) = n\pi$. If the signal is stationary and band-limited $f_{max}(t) = \text{const} = F_{max}$, Eq. (3) and (5) become equivalent. In case of non-constant $f_{max}(t)$ waveform of the reconstruction function $h_2(t, t_n)$ and the desired sampling instants t_n are determined by $f_{max}(t)$. Samples are spaced non-equidistantly and the mean sampling frequency can be less than it is required by Nyquist criterion, which, in this case, should be satisfied rather in local than in global sense.

2.2 Reconstruction algorithm

To reconstruct the non-uniformly sampled signal according to equation (1), the reconstruction procedure involves signal resampling to the equidistantly spaced sampling set $\{t_n\}$ with sampling period $\Delta t = t_n - t_{n-1} = \frac{1}{2F_{max}}$. The estimation of $\hat{s}(t_n)$ is possible according to the simple iterative algorithm [5] the idea of which is to interpolate the sampled band-limited signal $s(t)$ by the sum $\check{s}_{s(t_m)}(t) = \sum_m s(t_m)\psi_m$ and filter it in order to remove high frequencies. Piecewise linear interpolation, which is well suited to level-crossing samples, uses ψ_m consisting of the triangular functions

$$\psi_m(t) = \begin{cases} \frac{t - t_{m-1}}{t_m - t_{m-1}} & \text{for} \quad t_{m-1} \leq t < t_m, \\ \frac{t_{m+1} - t}{t_{m+1} - t_m} & \text{for} \quad t_m \leq t < t_{m+1}, \\ 0 & \text{elsewhere.} \end{cases} \tag{6}$$

It is proved [5] that if the maximum length of the gaps between the sampling instants $\tau_{max} \leq \frac{1}{2F_{max}}$, then every $s(t)$ can be reconstructed from the values $s(t_m)$ of an arbitrary τ_{max}-dense sampling set $\{t_m\}$ iteratively. The recovery algorithm can be written as:

$$\begin{aligned} &\hat{s}_0(t_n) = \check{s}_{s(t_m)}(t_n); \\ &\hat{s}_0(t) = C\left[\hat{s}_0(t_n)\right]; \\ &\hat{s}_i(t_n) = \hat{s}_{i-1}(t_n) + \check{s}_{(s - s_{i-1})(t_m)}(t_n); \\ &\hat{s}_i(t) = C\left[\hat{s}_i(t_n)\right], \end{aligned} \tag{7}$$

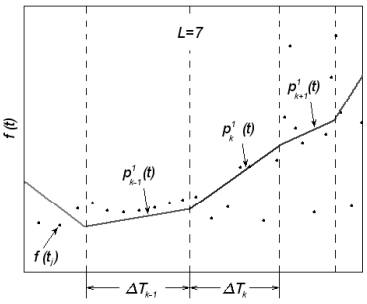


Figure 1: Piecewise polynomial $p_k^1(t)$ approximation.

where i indicates the number of iteration. The linear operator C denotes filtering as the convolution of samples $s(t_n)$ with impulse response $h_1(t, t_n)$ of the filter according to Eq. (4)

$$C\left[s(t_n)\right] = \sum_{n=0}^{N-1} s(t_n) h_1(t, t_n) \tag{8}$$

The sampling of non-stationary signal using level-crossing scheme does not ensure the satisfaction of the requirement $\tau_{max} \leq \frac{1}{2F_{max}}$. Direct application of the above described algorithm leads to a considerable reconstruction error, therefore two substantial enhancements are introduced to the algorithm - performing resampling to the non-equidistantly spaced values and the use of filter with impulse response $h_2(t, t_n)$ instead of classical $h_1(t, t_n)$. The resampling instants t_n are determined by $\Phi(t)$, which depends on $f_{max}(t)$, that in general case is not known in advance. To solve this problem, an algorithm is developed, which estimates the time-varying instantaneous maximum frequency using information about locations of level-crossings.

2.3 Estimation of instantaneous maximum frequency

The obvious ways to estimate the local bandwidth of the signal is by finding its time-frequency representation (TFR) using, for example, short-time Fourier transform, wavelet transform or Wigner-Ville distribution. These methods are developed for uniformly sampled signals, however, there are some modifications in order to find the TFR of non-uniformly sampled signals [7]. The use of such approach is time consuming, therefore a simpler method is considered that is based on empirical evaluations.

To estimate the function $\widehat{f}_{max}(t)$ from samples $s(t_m)$, starting with the initial index value $m = 0$ two pairs of successive level-crossing samples $s(t_{m'_j}) = s(t_{m'_j+1})$ and $s(t_{m''_j}) = s(t_{m''_j+1})$ are found such that $m''_j > m'_j$ and the difference $m''_j - m'_j$ is minimal. Thereafter the next two pairs are found considering that $m'_{j+1} = m''_j$. For each $j = 1, 2, \ldots$ the value $f(t_j)$ is calculated as

$$f(t_j) = \left(t_{m''_j} + t_{m''_j+1} - t_{m'_j} - t_{m'_j+1}\right)^{-1}, \tag{9}$$

where

$$t_j = \frac{1}{4}\left(t_{m''_j} + t_{m''_j+1} - t_{m'_j} - t_{m'_j+1}\right) \tag{10}$$

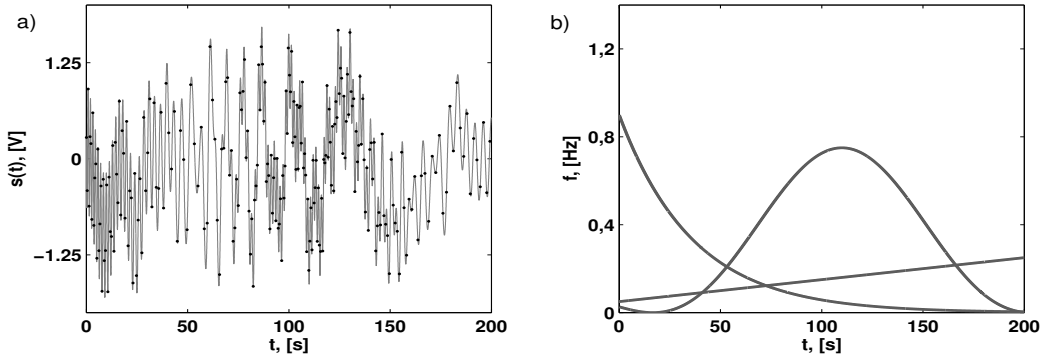


Figure 2: (a) Test signal sampled by $\Phi(t_n) = n\pi$ and (b) frequency traces of its components.

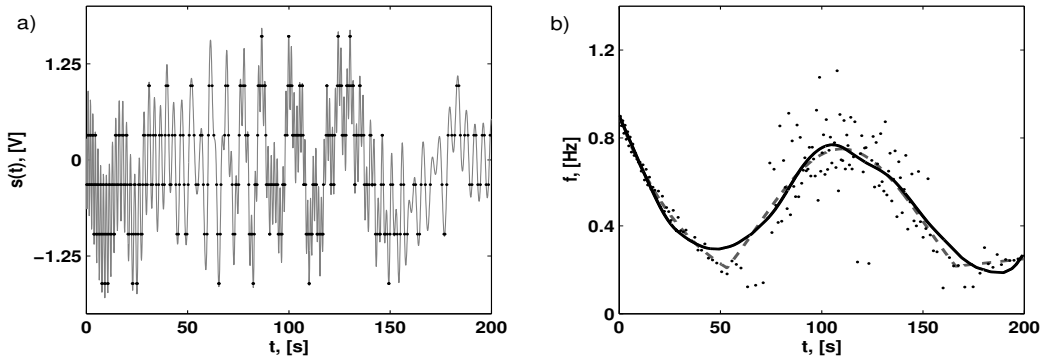


Figure 3: (a) Test signal sampled by level-crossings and (b) estimated instantaneous maximum frequency $\widehat{f}_{max}(t)$ as solid line, true instantaneous maximum frequency as dashed line and $f(t_j)$ as black points.

If a single sinusoid is sampled, then $f(t_j) = f(t_{j+1})$ for all j and it equals the frequency of the sinusoid. If the signal consists of more harmonics, then $f(t_j)$ for different j vary around the average value of $\bar{f} = \frac{1}{J}\sum_{j=1}^{J} f(t_j)$, where J is the total number of detected pairs within the observation time of the signal. Experiments show that $\bar{f}$ is close to the frequency of the highest component. Thus, the estimate of function of instantaneous maximum frequency $\widehat{f}_{max}(t)$ can be obtained by $\{f(t_j)\}$ approximation with piecewise polynomials $p_k^r(t)$ of order r. By choosing the number $L > 1$ the observation interval of signal is divided into subintervals

$$\Delta T_k : t \in [t_{k,1}; t_{k,2}]\,, \tag{11}$$

where $k = 0, 1, \ldots$ is the number of subinterval and

$$\begin{aligned} t_{k,1} &= \frac{t_{j=kL} + t_{j=kL+1}}{2}, \\ t_{k,2} &= \frac{t_{j=(k+1)L} + t_{j=(k+1)L+1}}{2} \end{aligned} \tag{12}$$

For each subinterval ΔT_k the coefficients $a_{k,r}, a_{k,r-1}, \ldots, a_{k,1}, a_{k,0}$ of polynomial $p_k^r(t) = a_{k,r}t^r + a_{k,r-1}t^{r-1} + \cdots + a_{k,1}t + a_{k,0}$ are found to ensure

$$\begin{aligned} &p_{k-1}^r(t_{k,1})^{(0)} = p_k^r(t_{k,1})^{(0)}\ ,\ p_k^r(t_{k,2})^{(0)} = p_{k+1}^r(t_{k,2})^{(0)} \\ &p_{k-1}^r(t_{k,1})^{(1)} = p_k^r(t_{k,1})^{(1)}\ ,\ p_k^r(t_{k,2})^{(1)} = p_{k+1}^r(t_{k,2})^{(1)} \\ &\qquad\qquad\vdots \\ &p_{k-1}^r(t_{k,1})^{(r)} = p_k^r(t_{k,1})^{(r)}\ ,\ p_k^r(t_{k,2})^{(r)} = p_{k+1}^r(t_{k,2})^{(r)} \end{aligned}$$

and the value of expression

$$\sum_{k=0}^{K-1}\sum_{j=kL+1}^{(k+1)L} [f(t_j) - p_k^r(t_j)]^2 = \min \tag{13}$$

is minimal. The denotation $(\ldots)^{(r)}$ means the derivative of order r and K is the total number of subintervals. After solving the minimization task using the method of least squares, the coefficients of polynomials $p_k^r(t)$ are obtained and the estimate of instantaneous maximum frequency

$$\widehat{f}_{max}(t) = p_k^r(t),\ \text{if } t_{k,1} \le t \le t_{k,2} \tag{14}$$

depends on the number L of samples $f(t_j)$ per subinterval. To reduce the dependency the final frequency estimate is obtained by averaging $\widehat{f}_{max}(t)$ calculated for different L values. The example of piecewise polynomial of order $r = 1$ approximation when $L = 7$ is shown in Fig. 1

3. Simulation results

The methods described in previous section are applied to reconstruct nonstationary signal from its nonuniform sam-

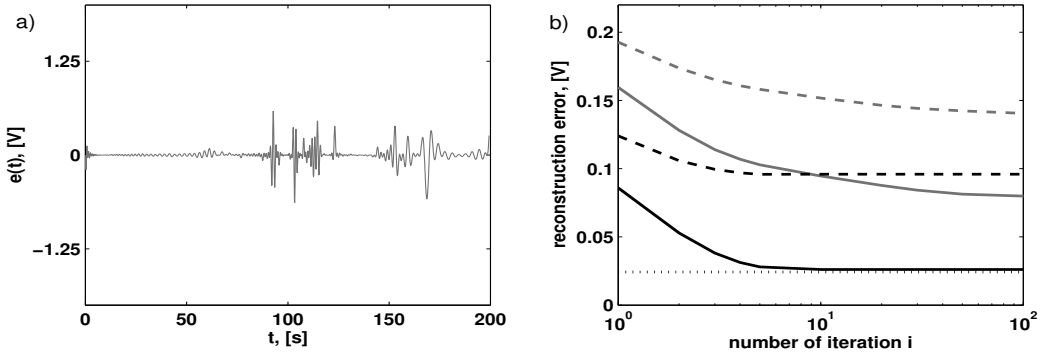


Figure 4: (a) The difference between original and recovered signal from its 349 level-crossing samples after 10 iterations and (b) reconstruction error (solid lines - reconstruction from level-crossings using $h_2(t,t_n)$, dashed lines - reconstruction from level-crossings using $h_1(t,t_n)$, dotted line - reconstruction from samples obtained by $\Phi(t_n)=n\pi$).

ples $s(t_n)$ obtained in two different ways. The first one is when $f_{max}(t)$ is given and sampling instants t_n satisfy $\Phi(t_n)=n\pi$ (Fig. 2). The second way is by level-crossing sampling and $f_{max}(t)$ is not known in advance (Fig. 3).
In the first case 239 nonequidistantly spaced samples were obtained during 200 seconds of the test signal, which consists of three sinusoids with constant amplitudes and time-varying frequencies as shown in Fig. 2b. As the reconstructed signal according to Eq. (4) using $h_2(t,t_n)$ differs insignificantly from the original one, it is not illustrated here. In order to obtain similar result in uniform sampling case, at least 360 samples would be required since the maximum frequency of the signal is $F_{max}=0.9$ Hz.
In the level-crossing sampling case 349 samples were captured using 6 quantization levels (Fig. 3a). To recover the signal the first task was to find the values $f(t_j)$ according to Eq. (9) in order to estimate the instantaneous maximum frequency (14). In Fig. 3b $f(t_j)$ are shown as black points, true $f_{max}(t)$ as dashed line and calculated $\widehat{f}_{max}(t)$ as solid line. The similarity between frequency traces is obvious. The second step was to recover the original signal according to Eq. (7) using level-crossing samples and estimated $\widehat{f}_{max}(t)$. The difference signal $e_i(t)=s(t)-s_i(t)$ after 10 iterations $i=10$ is illustrated in Fig. 4a. The reconstruction error $\sqrt{\frac{1}{T}\int_0^T e_i(t)^2dt}$ reduces as the number of iterations i increases. It is shown in Fig. 4b as a grey solid line. The grey dashed line corresponds to reconstruction error, when instead of time-varying bandwidth filter $h_2(t,t_n)$ the filter with constant cut-off frequency of $F_{max}=0.9$ Hz and impulse response $h_1(t,t_n)$ is used. In this case the achieved result is not so good as the reconstruction quality remains only in intervals, where the sampling density is sufficient. The reconstruction error can be reduced by decreasing the distance between quantization levels giving 437 level-crossing samples. It is shown in Fig. 4b as black solid and dashed lines. The dotted line corresponds to the first case when $f_{max}(t)$ is given and sampling instants t_n satisfy $\Phi(t_n)=n\pi$.

4. Conclusions

The proposed approach for non-stationary signal processing uses signal dependent techniques: level crossing sampling for data acquisition and filtering with time-varying bandwidth for signal reconstruction. The information carried by level-crossing samples is employed in two ways – time instants of samples are used to estimate the instantaneous maximum frequency of the signal, while the amplitude values of samples are used in reconstruction algorithm. The reconstruction procedure is based on the use of iterative filtering with time-varying bandwidth filter. The enhancement of classical signal reconstruction approach is made by introducing signal-dependent, "non-stationary" impulse response and resampling to the corresponding, nonuniform sampling set.
Speech signal processing can be quoted as one of the potential application areas of the proposed algorithm. The level-crossing sampling technique reduces the number of samples and leads to effective signal coding approaches.

References:

[1] P. Ellis. Extension of phase plane analysis to quantized systems. *IRE Transactions on Automatic Control*, 4(2):43–54, 1959.

[2] M. Miskowicz. Send-on-delta concept: An event-based data reporting strategy. *Sensors*, 6:49–63, 2006.

[3] E. Allier and G. Sicard. A new class of asynchronous a/d converters based on time quantization. In *Proc. of International Symposium on Asynchronous Circuits and Systems ASYNC'03*, pages 196–205, 2003.

[4] M. Greitans. Processing of non-stationary signal using level-crossing sampling. In *Proc. of the International Conference on Signal Processing and Multimedia Applications SIGMAP'06*, pages 170–177, 2006.

[5] H. G. Feichtinger and K. Grochening. Theory and practice of irregular sampling. 1994.

[6] M. Greitans and R. Shavelis. Speech sampling by level-crossing and its reconstruction using spline-based filtering. In *Proceedings of the 14th International Conference IWSSIP 2007*, pages 305–308, 2007.

[7] M. Greitans. Time-frequency representation based chirp-like signal analysis using multiple level crossings. In *Proceedings of the 15th European Signal Processing Conference EUSIPCO 2007*, 2007.

Optimal Characteristic of Optical Filter for White-Light Interferometry based on Sampling Theory

Hidemitsu Ogawa [(1)] and Akira Hirabayashi [(2)]

(1) Toray Engineering Co., Ltd., 1-45, Oe 1-chome, Otsu, Shiga, 520-2141, Japan.
(2) Yamaguchi University, 2-16-1, Tokiwadai, Ube City, Yamaguchi 755-8611, Japan.
hidemitsu-ogawa@kuramae.ne.jp, a-hira@yamaguchi-u.ac.jp

Abstract:

White-light interferometry is a technique of profiling surface topography of objects such as semiconductors, liquid crystal displays (LCDs), and so on. The world fastest surface profiling algorithm utilizes a generalized sampling theorem that reconstructs the squared-envelope function $r(z)$ directly from an infinite number of samples of the interferogram $f(z)$. In practical measurements, however, only a finite number of samples of the interferogram $g(z) = f(z) + C$ with a constant C are acquired by an interferometer. We have to estimate the constant C and to truncate the infinite series in the sampling theorem. In order to reduce both the truncation error and the estimation error for C, we devise an optimal characteristic of the optical filter installed in the interferometer in the sense that the second moment of the square of the interferogram is minimized. Simulation results confirm the effectiveness of the optimal characteristic of the optical filter.

1. Introduction

White-light interferometry is a technique of profiling surface topography of objects such as semiconductors, liquid crystal displays (LCDs), and so on. It is attractive because of its advantages including non-contact measurement and unlimited measurement range in principle [1, 2, 3, 5, 6, 8, 9]. From the viewpoint of sampling theory, white-light interferometry has the following two interesting features. First, a signal to be processed, a white-light interferogram, $f(z)$, is a bandpass signal. Second, a signal to be reconstructed from sampled values of $f(z)$ is not the interferogram itself, but its squared-envelope function $r(z)$. This type of sampling theorem is called a generalized sampling theorem [4, 10, 11].

The present authors also derived such a sampling theorem [9]. Based on the theorem, the world fastest surface profiling algorithm were proposed and installed in commercial systems [5]. The sampling theorem is expressed in a form of infinite series and uses samples of the interferogram $f(z)$. In practical measurements, however, only a finite number of samples of the interferogram $g(z) = f(z) + C$ with a constant C are acquired by an interferometer. Hence, in the algorithm, the constant C is estimated from the samples, and the infinite series is truncated with the number of samples. If both the truncation error and the estimation error for C were reduced, we can further improve the preciseness of the algorithm. For both error reductions, it is very effective for interferograms to have small side lobes. The waveform of interferograms can be controlled by an optical filter installed in the interferometer.

Hence, in this paper, we devise an optimal characteristic of the optical filter in the sense that the second moment of the square of the interferogram is minimized with a fixed band-width. We show that the optical characteristic is given by a sine curve which has a half of the period as the fixed band-width. We also show that we have a so-called uncertainty principle between the band-width and the second moment. Simulation results confirm the effectiveness of the optimal characteristic of the optical filter.

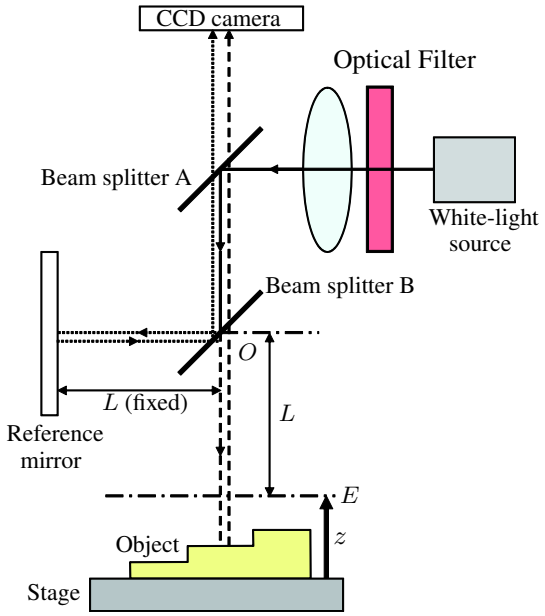


Figure 1: Basic setup of an optical system used for surface profiling by white-light interferometry.

2. Surface Profiling by White-Light Interferometry

Figure 1 shows a basic setup of an optical system used for surface profiling by white-light interferometry. It uses the Michelson interferometer. A beam from a white-light source passes through an optical filter. The beam is re-

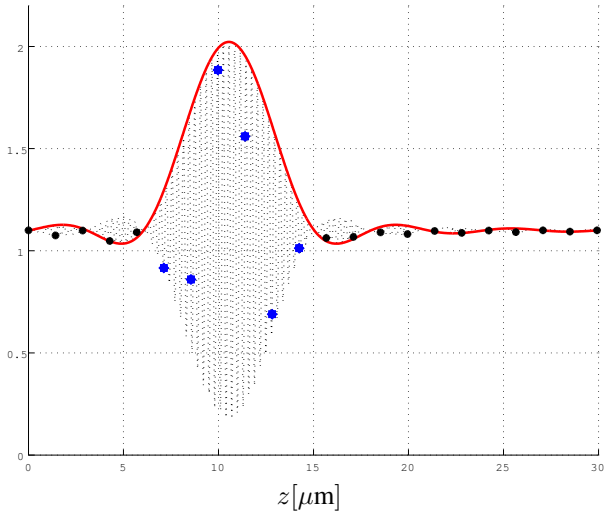


Figure 2: An example of a white-light interferogram $g(z)$ and its sampled values.

flected by the beam splitter A, and divided into two portions by the beam splitter B at the point O. One of the portions indicated by the dotted line is transmitted to a reference mirror, whose distance from the point O is L. The other portion indicated by the dashed line is transmitted to a surface of an object being observed. The height of the surface from the stage at the point P is denoted by z_p. E is a virtual plane whose distance from the point O is L. z is the distance of the plane E from the stage.

The two beams reflected by the object surface and the reference mirror are recombined and interfere. As the interferometer scans along the z-axis, the resultant beam intensity varies as is shown in Fig. 2 by the dotted line. It is called a white-light interferogram or simply an interferogram and denoted by $g(z) = f(z) + C$, where C is a constant. Its peak appears in the right side in Fig. 2 if the height z_p is high, while it appears in the left side if z_p is low. Hence, the maximum position of the interferogram provides the height z_p.

The intensity is observed by a charge-coupled device (CCD) video camera with a shutter speed of 1/1000 second. It has, for example, 512×480 detectors. Each of them corresponds to a point on the surface to be measured. Since the CCD camera outputs the intensity, for example, every $1/60$ second, we can utilize only discrete sampled values of the interferogram shown by '•' in Fig. 2. We have to estimate the maximum position of the interferogram from these sampled values.

It is known that the envelope function $m(z)$ shown by the solid line in Fig. 2, or its square $r(z)$, has the same peak as the interferogram and they are much smoother than the interferogram. Hence, usually these functions are used for detection of the peak instead of the interferogram. In this paper, we use the latter $r(z)$, which we call the squared-envelope function.

3. Sampling theorem for squared-envelope functions

Since the interferogram $f(z)$ is a bandpass signal, it can be reconstructed from its samples by using the sampling theorem for bandpass signals [7]. It is interesting that, since the squared-envelope function $r(z)$ is the sum of squares of $f(z)$ and its Hilbert transform, the squared-envelope function is also reconstructed from samples of $f(z)$, not those of $r(z)$. Indeed, the following result was established [9, 5]. The center wavelength and the bandwidth of the optical filter in Fig. 1 are denoted by λ_c and $2\lambda_b$, respectively. Let k_l and k_u be angular wavenumbers defined by

$$k_l = \frac{2\pi}{\lambda_c + \lambda_b}, \quad k_u = \frac{2\pi}{\lambda_c - \lambda_b}. \tag{1}$$

Two parameters $\omega_l = 2k_l$ and $\omega_u = 2k_u$ are also used.

Proposition 1 [5] (*Sampling theorem for squared-envelope functions*) *Let I be an integer such that*

$$0 \le I \le \frac{\omega_l}{\omega_u - \omega_l}, \tag{2}$$

and ω_b be any real number that satisfies

$$\begin{cases} \dfrac{\omega_u}{2} \le \omega_b & (I = 0), \\ \dfrac{\omega_u}{2(I+1)} \le \omega_b \le \dfrac{\omega_l}{2I} & (I \ne 0). \end{cases} \tag{3}$$

Let ω_c be a real number defined by

$$\omega_c = (2I+1)\omega_b. \tag{4}$$

Let Δ be a sampling interval given by

$$\Delta = \frac{\pi}{2\omega_b}, \tag{5}$$

and $\{z_n\}_{n=-\infty}^{\infty}$ be sample points defined by

$$z_n = n\Delta. \tag{6}$$

Then, it holds that

1. *When z is a sample point z_j,*

$$r(z_j) = \{f(z_j)\}^2 + \frac{4}{\pi^2}\left\{\sum_{n=-\infty}^{\infty} \frac{f(z_{j+2n+1})}{2n+1}\right\}^2. \tag{7}$$

2. *When z is not any sample point,*

$$r(z) = \frac{2\Delta^2}{\pi^2}\left[\left(1 - \cos\frac{\pi z}{\Delta}\right)\left\{\sum_{n=-\infty}^{\infty} \frac{f(z_{2n})}{z - z_{2n}}\right\}^2 + \left(1 + \cos\frac{\pi z}{\Delta}\right)\left\{\sum_{n=-\infty}^{\infty} \frac{f(z_{2n+1})}{z - z_{2n+1}}\right\}^2\right]. \tag{8}$$

To apply Proposition 1 for surface profiling, we have the following difficulties. In the proposition, an infinite number of sampled values $\{f(z_n)\}_{n=-\infty}^{\infty}$ of the interferogram $f(z)$ are used. In practical applications, however, only a finite number of sampled values $\{g(z_n)\}_{n=0}^{N-1}$ of the interferogram $g(z) = f(z) + C$ are available. Hence, we have to truncate the infinite series in Proposition 1 and approximate the sampled values $f(z_n)$ by $g(z_n) - \hat{C}$, where $\hat{C}$ is an estimate of C. For example, the average of $g(z_n)$ is used as $\hat{C}$. Now, we are suffered from the truncation error as well as the estimation error for $\hat{C}$. Both of these errors severely affect our final goal of precise estimation of z_p.

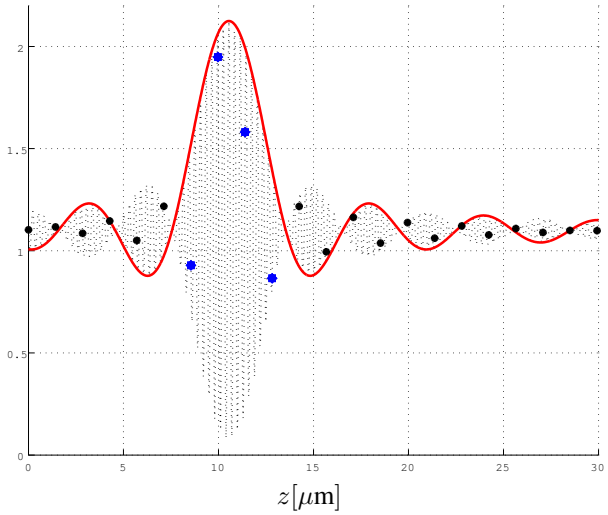


Figure 3: A white-light interferogram $g(z)$ when $\psi(k)$ is rectangular.

4. Optimal characteristics of optical filter

To reduce both of the errors, the following observation is crucial. As you can see in Fig. 2, only a few number of samples are located in the main lobe of $g(z)$ while the rest of them are in side lobes. The latter mostly vanishes once the constant C is estimated precisely. This implies that, the smaller the side lobes are, the smaller the truncation error is. Smaller side lobes also lead us to better estimations of C as shown experimentally in Section 5.

Fortunately, we can control the waveform of the interferogram by the optical filter in the interferometer. Let $a(k)$ be its characteristic in terms of an angular wavenumber k. The support of $a(k)$ is the interval $k_l < k < k_u$. Averaged attenuation rates of two beams along the dashed and the dotted lines in Figure 1 are denoted by $q_o(k)$ and $q_r(k)$, respectively. Let $\psi(k)$ be

$$\psi(k) = \begin{cases} 2\{a(k)\}^2 q_o(k) q_r(k) & (k > 0), \\ 0 & (k \le 0). \end{cases} \tag{9}$$

It is also supported on the same interval as $a(k)$:

$$\psi(k) = 0 \quad (k < k_l\,,\, k > k_u). \tag{10}$$

The function $\psi(k)$ is related to the interferogram $f(z)$ as

$$f(z) = \int_{k_l}^{k_u} \psi(k) \cos 2k(z - z_p)\, dk. \tag{11}$$

Equation (11) clearly shows that we can control $f(z)$ by $a(k)$ through $\psi(k)$.

To have smaller side lobes, we can easily arrive at the following idea: we design $\psi(k)$ so that it minimizes the second moment of the square of the interferogram $f(z)$:

$$J[\psi] = \int_{-\infty}^{\infty} (z - z_p)^2 \{f(z)\}^2 dz. \tag{12}$$

Now, we are at the point to show our main result in this paper. Let k_a be $(k_u - k_l)/2$.

Theorem 1 *Among second continuously differentiable functions $\psi(k) \in C^2[k_l, k_u]$ satisfying*

$$\psi(k) = 0 \quad (k \le k_l, k \ge k_u), \tag{13}$$
$$\psi(k) \ge 0 \quad (k_l < k < k_u), \tag{14}$$
$$\int_{k_l}^{k_u} \{\psi(k)\}^2 dk = 1, \tag{15}$$

$\psi(k)$ that minimizes the criterion $J[\psi]$ is given by

$$\psi(k) = \frac{1}{\sqrt{k_a}} \sin \frac{\pi(k - k_l)}{2k_a}. \tag{16}$$

The minimum value J_0 is given by

$$J_0 = \left(\frac{\pi}{2}\right)^3 \frac{1}{(2k_a)^2} = \frac{\pi\Delta^2}{2}. \tag{17}$$

The following two results are direct consequence of Theorem 1.

Corollary 1 *The optimal characteristic $a(k)$ under the criterion $J[\psi]$ is given by*

$$a(k) = \left(\frac{\sin \pi(k - k_l)/2k_a}{2\sqrt{k_a} q_o(k) q_r(k)} \right)^{1/2}. \tag{18}$$

Corollary 2 *The optimal waveform of the interferogram $f(z)$ is given by*

$$f(z) = m(z) \cos(k_u + k_l)(z - z_p), \tag{19}$$

where

$$m(z) = \frac{4\pi\sqrt{k_a} \cos 2k_a(z - z_p)}{\pi^2 - 16k_a^2(z - z_p)^2}. \tag{20}$$

The interferogram shown in Fig. 2 was the optimal one given by Eqs. (19) and (20) while that shown in Fig. 3 is generated from a rectangular $\psi(k)$ given by

$$\psi(k) = \begin{cases} 1/\sqrt{k_u - k_l} & (k_l < k < k_u), \\ 0 & (\text{otherwise}). \end{cases}$$

Though this $\psi(k)$ is not continuously second differentiable, the conditions (13) $\sim$ (15) are satisfied. In both figures, $\lambda_c = 600[nm]$ and $\lambda_b = 30[nm]$ were used. We can see that the side lobes in Fig. 2 are much smaller than those in Fig. 3. The sampling interval used in both figures is $\Delta = 1.425[\mu\mathrm{m}]$, which is the maximum among those satisfying Eqs. (2) $\sim$ (5). We have six samples in the main lobe in Fig. 2 while only four samples are located there in Fig. 3 (these samples are displayed by relatively large dots compared to samples in side lobes). In a nutshell, the optimal characteristic results in fewer samples in the small side lobes. This results in small errors on the truncation and the estimation of C, which we demonstrate in the next section through computer simulations.

Before proceeding simulations, let us make a final remark in this section.

Corollary 3 *Let σ^2 be the value of $J[\psi]$. Then, the following uncertainty principle holds:*

$$\sigma^2 (2k_a)^2 \ge \left(\frac{\pi}{2}\right)^3,$$
$$\frac{\sigma^2}{\Delta^2} \ge \frac{\pi}{2}.$$

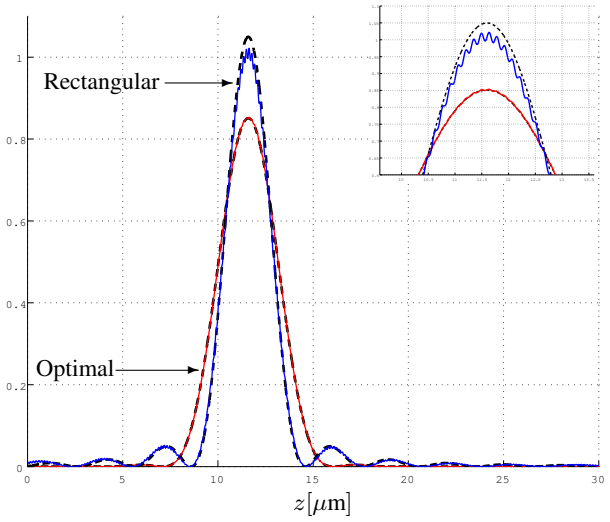


Figure 4: Squared-envelope functions (the dashed lines) and reconstructed functions (the solid lines) from samples of $g(z)$ for both of the optimal and the rectangular $\psi(k)$.

5. Simulations

We compare the optimal and the rectangular characteristics $\psi(k)$ by computer simulations. We first sample the interferograms $g(z)$ generated from both $\psi(k)$ with the sampling interval $\Delta = 1.425\mu$m. Then, the averages for each sample values are computed for the estimation of C. Finally, we reconstruct the squared-envelope functions $r(z)$ by using a finite number of $g(z_n) - \hat{C}$ instead of $f(z_n)$ in Proposition 1. The reconstructed functions are shown in Fig. 4 by the solid lines as well as the original squared-envelope functions by the dashed lines for both of the optimal and the rectangular $\psi(k)$. The small window in the top-right side shows the magnified image around the peak. We can see that the reconstructed function for the optimal $\psi(k)$ provides a much better result than that for the rectangular $\psi(k)$. We also notice that the latter oscillates severely.

The normalized truncation errors for the optimal and the rectangular $\psi(k)$ are 0.45% and 4.68%, respectively. The former is less than 10% of the latter. When $C = 1.10$, its estimation results are 1.10 and 1.06 for the optimal and the rectangular $\psi(k)$, respectively. Finally, errors for the estimation of z_p are 0.05μm and 0.06μm for the optimal and the rectangular $\psi(k)$, respectively. Even though the difference is not so significant, the oscillation of the reconstructed squared-envelope function for the rectangular $\psi(k)$ may cause difficulties for fast search of the maximum position.

We repeated the same simulations for thirty two values of z_p from 10μm to 20μm. Then, averages of estimation errors were 0.0496μm and 0.0541μm for the optimal and the rectangular, respectively. They are almost the same value. However, the averages of truncation errors were 0.35% and 4.67% for the optimal and the rectangular $\psi(k)$, respectively. The former is less than 7% of the latter. These results show the effectiveness of the optimal characteristics of the optical filter.

6. Conclusion

In this paper, we devised an optimal characteristic of the optical filter that minimizes the second moment of the square of the interferogram so that both of the truncation error and the estimation error for the constant in the interferogram are reduced. We showed that the optimal characteristic is given by a sine curve which has a half of the period as the band-width of the optical filter. Simulation results showed that the truncation error for the optimal characteristic is less than 7% of that for the rectangular one. The estimation error of the constant for the optimal characteristic was also smaller than the rectangular one. Even though the difference on the estimation error of the maximum position was not so significant, reconstructed functions for the optimal characteristic was much smoother than those for the rectangular one. These results showed the effectiveness of the optimal characteristic. Our future tasks include to produce a prototype of the optical filter with the optimal characteristic.

References:

[1] P.J. Caber. Interferometric profiler for rough surfaces. *Applied Optics*, 32(19):3438–3441, 1993.

[2] S.S.C. Chim and G.S. Kino. Three-dimensional image realization in interference microscopy. *Applied Optics*, 31(14):2550–2553, 1992.

[3] P. de Groot and L. Deck. Surface profiling by analysis of white-light interferograms in the spatial frequency domain. *Journal of Modern Optics*, 42(2):389–401, 1995.

[4] O.D. Grace and S.P. Pitt. Sampling and interpolation of bandlimited signals by quadrature methods. *The Journal of the Acoustical Society of America*, 48(6):1311–1318, 1969.

[5] A. Hirabayashi, H. Ogawa, and K. Kitagawa. Fast surface profiler by white-light interferometry by use of a new algorithm based on sampling theory. *Applied Optics*, 41(23):4876–4883, 2002.

[6] G.S. Kino and S.S.C. Chim. Mirau correlation microscope. *Applied Optics*, 29(26):3775–3783, 1990.

[7] A. Kohlenberg. Exact interpolation of band-limited functions. *Journal of Applied Physics*, 24:1432–1436, 1953.

[8] K.G. Larkin. Efficient nonlinear algorithm for envelope detection in white light interferometry. *Journal of Optical Society of America*, 13(4):832–843, 1996.

[9] H. Ogawa and A. Hirabayashi. Sampling theory in white-light interferomtery. *Sampling Theory in Signal and Image Processing*, 1(2):87–116, 2002.

[10] D.W. Rice and K.H. Wu. Quadrature sampling with high dynamic range. *IEEE Transactions on Aerospace and Electronic Systems*, AES-18(4):736–739, 1982.

[11] W.M. Waters and B.R. Jarrett. Bandpass signal sampling and coherent detection. *IEEE Transactions on Aerospace and Electronic Systems*, AES-18(4):731–736, 1982.

SAMPTA'09

Poster Sessions

Continuous Fast Fourier Sampling

Praveen K. Yenduri[(1)] and Anna C. Gilbert[(2)]

(1) University of Michigan, 4438 EECS building, Ann Arbor, MI 48109, USA.
(2) University of Michigan, 2074 East Hall, Ann Arbor, MI 48109, USA.
ypkumar@umich.edu, annacg@umich.edu

Abstract:

Fourier sampling algorithms exploit the spectral sparsity of a signal to reconstruct it quickly from a small number of samples. In these algorithms, the sampling rate is sub-Nyquist and the time to reconstruct the dominate frequencies depends on the type of algorithm—some scale with the number of tones found and others with the length of the signal. The Ann Arbor Fast Fourier Transform (AAFFT) scales with the number of desired tones. It approximates the DFT of a spectrally sparse digital signal on a fixed block by taking a small number of structured random samples. Unfortunately, to acquire spectral information on a particular block of interest, the samples acquired must be appropriately correlated for that block. In other words, the sampling pattern, though random, depends on the block of interest. When blocks of interest overlap significantly, the union of the sampling patterns may not be an optimal one (it might not be sub-Nyquist anymore). Unlike the much slower algorithms, the sampling pattern does not accommodate an arbitrary block position. We propose a new sampling procedure called Continuous Fast Fourier Sampling which allows us to continuously sample the signal at a sub-Nyquist rate and then apply AAFFT on any arbitrary block. Thus, we have a highly resource-efficient continuous Fourier sampling algorithm.

1. Introduction

Let x be a discrete time signal of length n which is sparse or compressible in the frequency domain but the exact frequency content depends on time. We consider the problem of computing the frequency content present in different blocks of the signal in a resource efficient manner. This problem arises in many applications such as *cognitive radio* [2] where a wireless node alters its transmission or reception parameters based on active monitoring of radio frequency spectrum at various times. Another application is *incoherent demodulation* of communication signals [3] such as FSK, MSK, OOK, etc., where the computed frequency spectrum at different times represents the message being transmitted itself.

There are several Fourier sampling algorithms [1, 8, 9] with low sampling costs that reconstruct the entire spectrum of a sampled signal. These algorithms make use of a uniformly random (not structured) sample set for computations thus allowing us to compute frequencies in any arbitrary block of interest from the signal. However, the time to reconstruct the spectrum is superlinear in signal's size and hence are slow and inappropriate for the applications involving large signal sizes or bandwidths where just a few frequencies are of interest. Instead, we consider a sub-linear time computational method called the AAFFT (Ann Arbor Fast Fourier Transform) described in [4].

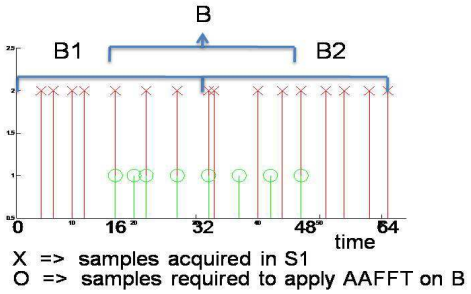


Figure 1: Figure showing the samples acquired in $S1$ and the samples required to apply AAFFT on $B = [16, 47]$.

Let y be a fixed block of interest of length N in the discrete time signal x. Since x is sparse in frequency domain, it can be assumed that y has only m dominant digital frequencies, where $m \ll N$. The AAFFT algorithm takes a small number of (correlated) random samples from the block of interest and produces an approximation of its DFT (identifies dominant tones), using time and storage $m\text{poly}(\log(N))$. If we are interested in a windowed Fourier analysis of x over windows of length N, a straightforward approach towards solving our problem using AAFFT is to divide the signal x into consecutive non-overlapping blocks of length N, generate appropriately correlated sampling patterns for each block, acquire the samples and then apply AAFFT on each block. Let us call this sample set $S1$. Unfortunately, $S1$ does not accommodate arbitrary block positions. For example, consider samples acquired in $S1$ from two consecutive blocks $B1$ and $B2$. Lets say we are now interested in block B which consists of second half of $B1$ and first half of $B2$ (see Figure (1)). However AAFFT cannot be applied on B since the samples acquired from B will not be appropriately structured for its application. This is illustrated in Figure (1) for a simple case of $N = 32$ with a dummy y-axis and a few samples plotted for clarity.

We propose a new sampling procedure called the Continuous Fast Fourier Sampling that allows us to continu-

ously sample the signal (as opposed to division into discrete blocks) at a sub-nyquist rate and then apply AAFFT on any arbitrary block of interest. The article describes the algorithm in detail in Section (3.2), proves its correctness in Section (3.3), followed by a few numerical experiments and results in Section (3).

2. The Fourier Sampling Algorithm (AAFFT)

The Fourier Sampling algorithm is predicated upon non-evenly spaced samples unlike many traditional spectral estimation techniques [6, 7] and uses a highly nonlinear reconstruction method that is divided into two stages, *frequency identification* and *coefficient estimation*, each of which includes multiple repetitions of basic subroutines. A detailed description of the implementation of AAFFT is available in [5].

Frequency Identification consists of two steps, dominant frequency isolation and identification. Isolation is carried out by a two-stage process: (i) pseudo random permutations of the spectrum, followed by (ii) the application of a filter bank with $K = O(m)$ bands, where m = number of tones (dominant spikes) in the signal. With high probability, a significant fraction of the dominant tones fall into individual bands, isolating each tone from the others and this probability can be increased with additional repetitions. Note that all the above is carried out *conceptually* in the frequency domain but instantiated in the time domain. That is, we sample the permuted and filtered signal in the time domain. To carry out the computations, the algorithm uses signal samples at time points indexed by $P(t,\sigma) = \{(t + q\sigma) \mod N, q = 0, 1, .., K-1\}$, where (t,σ) is randomly chosen for each repetition. The identification stage performs group testing to determine the dominant frequency value in each of the K outputs of the filterbank. This stage uses the samples indexed at arithmetic progressions $P(t^b,\sigma)$ formed from each element of the geometric progression $t^b = t + \frac{N}{2^{b+1}}$, $b = 0, 1, .., \log_2(N/2)$. The *estimation* stage uses the random sampling similar to the isolation stage for coefficient estimation of each of the dominant frequencies identified.

Note that although the (t,σ) pair is chosen randomly in each repetition, the samples that result from each pair are highly structured. Let $A_1 = \{(t,\sigma)\}$ used in the *frequency identification* stage and similarly let A_2 be defined for the *estimation* stage. These two sets define a sampling pattern.

3. Continuous Fast Fourier Sampling

3.1 Sample set construction

Let n be the length of signal x which has m dominant tones that vary over time. Let the block length be N. Let $K = O(m)$ and $\alpha = \log_2(N)$. Let (t,σ) be a fixed pair in A_1 or A_2. Define a sequence of time points $t(0) = t$, $t(j) = (t(j-1) + Q(j-1)\sigma) \mathrm{mod} N$ for $j = 1, .., J$, where $Q(j-1)$ = smallest integer such that $t(j-1)+Q(j-1)\sigma \geq N$ and $J = \lceil \frac{K\sigma}{N} \rceil$. We call $t(j)$ the "N-wraparound" of $t(j-1)$. Figure (2) illustrates the calculation of a N-wraparound. The choice of J is such that the theorems in Section (3.3) hold. For $j = 1, .., J$, denote by I_j the arithmetic progression formed by $(t(j),\sigma)$,

$$I_j = \{t(j) + q\sigma, \forall q \geq 0 : t(j) + q\sigma \leq n\} \qquad (1)$$

Now, consider the geometric progression $t^b = t + \frac{N}{2^{b+1}}$ for all $b = 0, 1, .., \alpha - 1$. For each b, $\left(t + \frac{N}{2^{b+1}}, \sigma\right)$ is treated as another (t,σ) pair and the sequence $t^b(j)$ and the corresponding progressions I_j^b can be defined.

Do all the above, for each pair (t_ℓ, σ_ℓ) in A_1 and A_2 and denote the arithmetic progressions produced, by $I_{\ell,j}$, for $j = 1, .., J_\ell$. Define the union of all such arithmetic progressions as $I_\ell = \bigcup_{j=0}^{J_\ell} I_{\ell,j}$. Similarly define $I_\ell^b = \bigcup_{j=0}^{J_\ell} I_{\ell,j}^b$ for $b = 0, .., \alpha - 1$. Now define $I_\ell^B = \bigcup_{b=0}^{\alpha-1} I_\ell^b$. Finally define

$$I(A_1, A_2) = \left(\bigcup_{A_1}(I_\ell \cup I_\ell^B)\right) \cup \left(\bigcup_{A_2} I_\ell\right) \qquad (2)$$

Given a set of indices I, we denote by $S^x(I)$ the set of samples from signal x indexed by I.

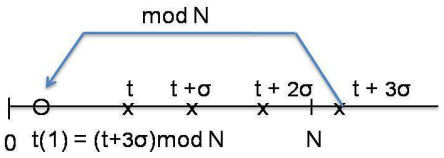


Figure 2: Calculation of N-Wraparound $t(1)$ from t.

3.2 The CFFS Algorithm

Preprocessing:
INPUT: N // Block length (1) Sample-set generation : Choose A_1 and A_2 as defined and compute $I(A_1, A_2)$ (as in Equation (2)). **OUTPUT:** $I(A_1, A_2)$ // Index set
Sample Acquisition
INPUT: $I(A_1, A_2)$, x (2) sample signal x at I and obtain samples $S^x(I)$. **OUTPUT:** $S^x(I)$
Reconstruction
INPUT: $S^x(I)$, (n_1, n_2) // boundary indices of an arbitrary block y of length N from signal x (3) calculate A_1', A_2' (depend on (n_1, n_2), defined in Section (3.3)) and extract $S^y(I(A_1', A_2')) \subset S^x(I)$. (4) apply AAFFT on the sample-set $S^y(I(A_1', A_2'))$ **OUTPUT:**top m frequencies of x in block $y = x[n_1, n_2]$

3.3 Proof of Correctness of CFFS

The arbitrary block y has boundaries (n_1, n_2). To generate samples from this block, we define new sets A_1' and A_2' as follows. For every (t,σ) in A_1 and A_2, let

i be the smallest integer such that $t + i\sigma > n_1$. Define $t' = (t + i\sigma) \text{mod} n_1$. Note that t' is simply the n_1-wraparound of t. Put $A'_1 = \{(t', \sigma) : (t, \sigma) \in A_1\}$ and similarly A'_2. Note that A'_1 and A'_2 are still random since A_1 and A_2 were chosen randomly. To apply AAFFT on block y we can now use samples of y indexed by the sampling pattern defined (as in Section (2.)) from A'_1 and A'_2. The following theorems together show that the required samples of y are available in $S^x(I(A_1, A_2))$.

Theorem 1 *For sets A'_1 and A'_2 as defined above, $S^y(I(A'_1, A'_2)) \subset S^x(I(A_1, A_2))$.*

Theorem 2 *AAFFT can be applied on the sample-set $S^y(I(A'_1, A'_2))$, i.e. the index set $I(A'_1, A'_2)$ has the required structure explained in Section (2.).*

Rather than giving detailed proofs, we prove a proposition that lies at the heart of the two theorems.

Proposition 3 *For every (t', σ) in A'_1 or A'_2, $S^y(P(t', \sigma)) \subset S^x(I(A_1, A_2))$.*

Proof: Let (t, σ) be the pair in A_1 or A_2 from which (t', σ) was obtained. We will prove that the arithmetic progressions I_j formed by the sequence of wraparounds $t(j)$,$j = 1, .., J$ as defined in Section (3.1), induce mod-N arithmetic in the progression $P(t', \sigma)$ (P as defined in Section (2.)). Consider the first few terms in $P(t', \sigma)$, till $(t' + (q_0 - 1)\sigma)$ mod N where q_0 is the smallest integer such that $(t' + q_0\sigma) \geq N$. From definition of t' observe that $t' = (t + i\sigma - n_1)$. so $y(t') = x(n_1 + t') = x(t + \sigma) \in S^x(I_0)$, where I_0 is defined in Equation (1). Similarly it is easy to see that the first q_0 terms in $S^y(P(t', \sigma))$ are contained in $S^x(I_0)$. Now call the next term $(t' + q_0\sigma)$ mod $N = t'(1)$. Observe that $t'(1) = t' + \sigma \left\lceil \frac{N - t'}{\sigma} \right\rceil - N$. Similarly observe that $t(1) = t + \sigma \left\lceil \frac{N-t}{\sigma} \right\rceil - N$. Now, Substituting $t' = (t + i\sigma - n_1)$ in the expression for $t'(1)$ we get, $t'(1) = t + i\sigma - n_1 + \sigma \left\lceil \frac{N - t + n_1 - i\sigma}{\sigma} \right\rceil - N = t + i\sigma - n_1 + \sigma \left\lceil \frac{N-t}{\sigma} \right\rceil + d\sigma - N = t(1) + (i + d)\sigma - n_1$, for an appropriately defined d, which can be shown to be positive. So $y((t' + q_0\sigma) \text{ mod } N) = y(t'(1)) = x(t(1) + (i + d)\sigma) \in S^x(I_1)$, where again I_1 is defined in Equation (1). Let q_1 be the smallest integer such that $(t'(1) + q_1\sigma) \geq N$. Now it is easy to see that the next q_1 terms in $S^y(P(t', \sigma))$ are contained in $S^x(I_1)$. Repeat this until all the terms in $P(t', \sigma)$ are covered. ∎

Proposition 4 *On average, the storage requirement of CFFS algorithm is $O(\frac{n}{N} m \log^{O(1)} N)$, which is of the same order as a straightforward, fixed boundary sample set for AAFFT.*

4. Results and Discussion

The Continuous Fast Fourier Sampling algorithm has been implemented and tested in various settings. In particular, we performed following three experiments.
First, we consider a model problem for communication devices which use frequency-hopping modulation schemes. The signal we want to reconstruct has two tones that

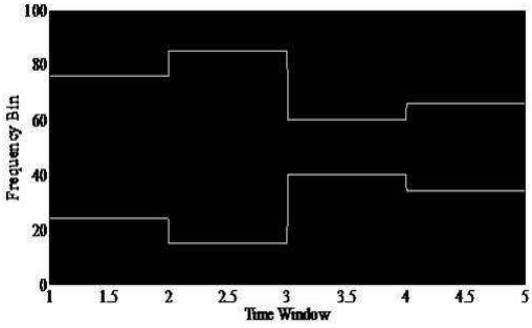


Figure 3: The Sparsogram for a synthetic frequency-hopping signal consisting of two tones, as computed by AAFFT ($S1$) and by CFFS.

change at regular intervals. We apply both the straightforward AAFFT on $S1$ and CFFS to identify the location of the tones. Figure (3) shows the obtained *sparsogram* which is a time-frequency plot that displays only the dominant frequencies in the signal. We get the same sparsogram in both cases, as expected. For $N = 2^{20}$, $S1$ samples about 0.94% of the signal whereas CFFS samples about 1.06% of the signal, which is only very slightly larger than $S1$. This experiment demonstrates the efficiency and similarity of the two methods and supports the proposition made in Section (3.3).

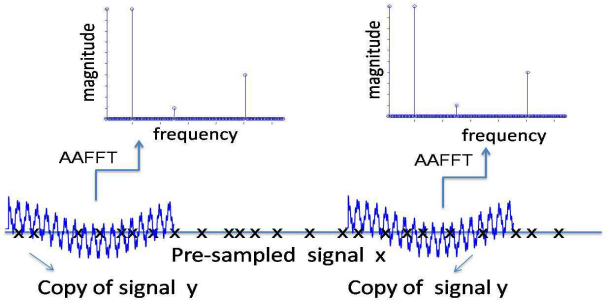


Figure 4: Applying CFFS to different blocks of signal x.

While $S1$-AAFFT cannot be applied to compute the dominant tones in any arbitrary block, the CFFS has no such limitation. This is demonstrated in the next experiment as follows. Let y be a signal of length $N = 2^{20}$, with $m = 4$ known dominant frequencies. Let x be an arbitrary signal of length n with $N \ll n$. Now let $x[n_1, n_2]$ be an arbitrary block of interest of length N. Set $x(n_1 + q) = y(q)$, for $q = 0, 1, \ldots, N - 1$. Thus we have placed a copy of the known signal y in the block of interest. The CFFS was then applied and the four dominant frequencies in the block of interest were computed. The obtained values for frequencies and their coefficients match closely with those of the signal y and satisfy the error guarantees of AAFFT. The whole experiment was repeated with different values for n_1 (and corresponding $n_2 = n_1 + N - 1$) and the same results were obtained. Figure (4) shows the sketch of a signal x, pre-sampled in a predetermined manner (according to CFFS), with copies of y placed at arbitrary positions. Application of AAFFT to any block with copy of y gives the same results thus demonstrating the correctness of CFFS.

In the final experiment, we consider the frequency hopping signal from the first experiment. Let the block size be $N = 2^{17}$ with unknown block boundaries. Let f_1 and f_2 be the respective frequencies in two adjacent blocks ($f1$ in the left block). We consider the problem of finding the block boundary using CFFS with an analysis window of size N. The center of the window can be varied and a binary search can be performed for the block boundary in the following manner. If the center is to the left of the actual boundary, then the coefficient of $f1$ produced by AAFFT will be higher than that of the $f2$. This indicates that the center has to be moved to the right from its current position. Also the search is not strictly binary since the amount by which $f1$ coefficient is higher than $f2$ can be used to shift the center of the window to the right by an equivalent amount. This step can be iterated a few times to make the center converge to the actual block boundary. We express the error as the distance to the true boundary and determine what percentage of the block this distance is. Table (1) displays the error and how the error increases with decreasing SNR. Note that even in the case

SNR(dB)	%Error	SNR(dB)	%Error
no noise	0.39	6	0.78
10	0.58	4	0.79
8	0.70	2	1.56

Table 1: Percentage error in boundary identification.

of no noise there is some inherent ambiguity in the identification of block boundary. This uncertainty is caused by two factors. First, when the analysis window has portions of both the $f1$-block and $f2$-block, the net signal is no longer sparse due to a sudden change in frequency and has a slowly decaying spectrum. With $m = 2$ the AFFT guarantees that the error made in signal approximation is about as much as the error in optimal 2-term approximation [5]. Hence a slowly decaying spectrum implies more error in the approximation. A second and more important factor is the number of samples actually acquired from the region of uncertainty around the block boundary. From the entire block, CFFS acquires about 8% samples from the $N = 2^{17}$ present. Assuming these samples are uniformly distributed (which is not true for CFFS), the number of samples present in the region of uncertainty (0.4%) is about 40. In practice, CFFS contains even fewer samples in the uncertainty region (about 30 on average). In terms of samples actually acquired in CFFS, the boundary estimation is off by only a few samples and hence is negligible, as it does not affect the computations. This will be true for any sparse sampling method like CFFS. Furthermore, if the uncertainty were to be reduced to 0.3% say, the boundary identification would improve by only about 6 samples on average, which again is negligible. Hence the boundary identification through the above method is accurate enough for all practical purposes.

5. Conclusions and Future Work

We described and proved a sub-linear time sparse Fourier sampling algorithm called the CFFS which along with AAFFT can be applied to compute the frequency content of sparse digital signals at any point of time. Once the block length N is selected, a sub-nyquist sampling pattern can be pre-determined and the samples can be acquired from the signal (during the runtime if required). The AAFFT can be applied to the samples corresponding to any block of length N of the signal and the dominant frequencies in that block and their coefficients can be computed in sublinear time. The algorithm requires the block length N to be fixed beforehand. Designing or extending the algorithm to work for different values of N can be considered. Adapting the algorithm to further reduce the computational complexity by using known side information about the signal can also be considered. The algorithm is also highly parallelizable and can be adapted for hardware applications. Also, we may be able to extend this sample set generation to the deterministic sampling algorithm described in [10].

Acknowledgements

The authors have been partially supported by DARPA ONR N66001-06-1-2011. ACG is an Alfred P. Sloan Fellow.

References:

[1] E.Candes, J.Romberg and T.Tao, Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans.Inform.Theory*, 52:489–509, 2006.

[2] I.F.Akyildiz, W.Y.Lee, M.C.Vuran, and S.Mohanty, Next generation dynamic spectrum access cognitive radio wireless networks: A survey, *Computer Networks Journal (Elsevier)*,50: 2127–2159, Sep 2006.

[3] Simon Haykin, *Communication systems*. Fourth Edition, John Wiley and Sons, 2005.

[4] A.C.Gilbert, S.Muthukrishnan, and M.J.Strauss, Improved time bounds for near-optimal sparse Fourier representations. *Proc. SPIE Wavelets XI*, 59141(A):1–15, 2005.

[5] A.C.Gilbert, M.J.Strauss, and J.A.Tropp. A Tutorial on Fast Fourier Sampling *IEEE Signal Processing Magazine*, 25(2):57–66, March 2008.

[6] G.K. Smith and D.M. Hawkins, Robust frequency estimation using elemental sets. *J.Comput.Graph.Stat*, 9(1):196–214, 2000.

[7] G. Harikumar and Y. Bresler, FIR perfect signal reconstruction from multiple convolutions: minimum deconvolver orders. *IEEE Trans.Signal Processing*, 46(1):215–218, 1998.

[8] G. Cormode and S. Muthukrishnan. Combinatorial algorithms for compressed sensing. *Proc.2006 IEEE Int.Conf.Information Sciences Systems*, 230–294, April 2006.

[9] A.C.Gilbert, M.J.Strauss, J.A.Tropp, and R.Vershynin, One sketch for all: Fast algorithms for Compressed Sensing. *Proc. 39th ACM Symposium on Theory of Computing*, 237–246, June 2007.

[10] M.A.Iwen, A deterministic sub-linear time sparse Fourier algorithm via non-adaptive compressed sensing methods. *SODA '08*, 20–29, Jan 2008.

Double Dirichelet Averages and Complex B-Splines

Peter Massopust [1,2]

(1) Institute for Biomathematics and Biometry, Helmholtz Zentrum München, Neuberberg, Germany.
(2) Center of Mathematics, Technische Universität München, Garching, Germany.
massopust@ma.tum.de

Abstract:

A relation between double Dirichlet averages and multivariate complex B-splines is presented. Based on this reationship, a formula for the computation of certain moments of multivariate complex B-splines is derived.

1. Introduction

Recently, a new class of B-splines with complex order z, $\operatorname{Re} z > 1$, was introduced in [4]. It was shown that complex B-splines generate a multiresolution analysis of $L^2(\mathbb{R})$. Unlike the classical cardinal B-splines, complex B-splines B_z possess an additional modulation and phase factor in the frequency domain:

$$\widehat{B}_z(\omega) = \widehat{B}_{\operatorname{Re} z}(\omega)\, e^{i \operatorname{Im} z \ln|\Omega(\omega)|}\, e^{-\operatorname{Im} z \arg \Omega(\omega)},$$

where $\Omega(\omega) := (1 - e^{-i\omega})/(i\omega)$. The existence of these two factors allows the extraction of additional information from sampled data and the manipulation of images.

In [6] and [9], some further properties of complex B-splines were investigated. In particular, connections between complex derivatives of Riemann-Liouville or Weyl type and Dirichlet averages were exhibited. Whereas in [6] the emphasis was on univariate complex B-splines and their applications to statistical processes, multivariate complex B-splines were defined in [9] using a well-known geometric formula for classical multivariate B-splines [7, 10]. It was also shown that Dirichlet averages are especially well-suited to explore the properties of multivariate complex B-splines. Using Dirichlet averages, several classical multivariate B-spline identities were generalized to the complex setting. There also exist interesting relationships between complex B-splines, Dirichlet averages and difference operators, several of which are highlighted in [5].

This short paper presents a generalization of some results found in [3, 12] to complex B-splines. For this purpose, the concept of double Dirichlet average [1] was introduced and its definition extended via projective limits to an infinite-dimensional setting suitable for complex B-splines. Moments of complex B-splines are defined and a formula for their computation in terms of a special double Dirichlet average presented.

2. Complex B-Splines

Let $n \in \mathbb{N}$ and let $\triangle^n$ denote the standard n-simplex in $\mathbb{R}^{n+1}$:

$$\triangle^n := \left\{ u := (u_0, \ldots, u_n) \in \mathbb{R}^{n+1} \;\middle|\; u_j \geq 0;\; j = 0, 1, \ldots, n; \sum_{j=0}^{n} u_j = 1 \right\}.$$

The extension of $\triangle^n$ to infinite dimensions is done via projective limits. The resulting infinite-dimensional standard simplex is given by

$$\triangle^\infty := \left\{ u := (u_j)_j \in (\mathbb{R}_0^+)^{\mathbb{N}_0} \;\middle|\; \sum_{j=0}^{\infty} u_j = 1 \right\},$$

and endowed with the topology of pointwise convergence, i.e., the weak$*$-topology. We denote by $\mu_b = \varprojlim \mu_b^n$ the projective limit of *Dirichlet measures* μ_b^n on the n-dimensional standard simplex $\triangle^n$ with density

$$\frac{\Gamma(b_0)\cdots\Gamma(b_n)}{\Gamma(b_0 + \cdots + b_n)}\, u_0^{b_0 - 1}\, u_1^{b_1 - 1} \cdots u_n^{b_n - 1}. \qquad (1)$$

Here, $\Gamma : \mathbb{C}\backslash\mathbb{Z}_0^- \to \mathbb{C}$ denotes the Euler Gamma function. Let $\mathbb{R}^+ := \{x \in \mathbb{R} \mid x > 0\}$ and let $\mathbb{C}_+ := \{z \in \mathbb{C} \mid \operatorname{Re} z > 0\}$.

Definition 1 ([6])**.** Given a weight vector $b \in \mathbb{C}_+^{\mathbb{N}_0}$ and an increasing knot sequence $\tau := \{\tau_k\}_k \in \mathbb{R}^{\mathbb{N}_0}$ with the property that $\lim_{k\to\infty} \sqrt[k]{\tau_k} \leq \varrho$, for some $\varrho \in [0, e)$, a complex B-spline $B_z(\bullet \mid b; \tau)$ of order z, $\operatorname{Re} z > 1$, with weight vector b and knot sequence τ is a function satisfying

$$\int_{\mathbb{R}} B_z(t \mid b; \tau) g^{(z)}(t)\, dt = \int_{\triangle^\infty} g^{(z)}(\tau \cdot u)\, d\mu_b(u) \qquad (2)$$

for all $g \in \mathscr{S}(\mathbb{R})$.

Here, $\mathscr{S}(\mathbb{R})$ denotes the space of Schwartz functions on $\mathbb{R}$, and $\tau \cdot u = \sum_{k\in\mathbb{N}_0} \tau_k u_k$ for $u = \{u_k\}_{k\in\mathbb{N}_0} \in \triangle^\infty$. In addition, we used the Weyl or Riemann-Liouville fractional derivative [8, 11, 13] of complex order z, $\operatorname{Re} z > 0$, $W^z : \mathscr{S}(\mathbb{R}) \to \mathscr{S}(\mathbb{R})$, defined by

$$(W^z f)(x) := \frac{(-1)^n}{\Gamma(\nu)} \frac{d^n}{dx^n} \int_x^\infty (t - x)^{\nu - 1} f(t)\, dt,$$

with $n = \lceil \operatorname{Re} z \rceil$, and $\nu = n - z$. Here $\lceil \cdot \rceil : \mathbb{R} \to \mathbb{Z}$, $x \mapsto \min\{n \in \mathbb{Z} \mid n \geq x\}$, denotes the *ceiling function*. To simplify notation, we write $f^{(z)}$ for $W^z f$

It is easy to show that the univariate complex B-spline $B_z(t \mid b; \tau)$ is an element of $L^2(\mathbb{R})$ [5].

Remark 2. For finite $\tau = \tau(n)$ and $b = b(n)$ and $z := n \in \mathbb{N}$, (2) defines also *Dirichlet splines* if g is chosen in $C^n(\mathbb{R})$. For, Dirichlet splines $D_n(\cdot \,|\, b; \tau)$ of order n are defined as those functions for which

$$\int_{\mathbb{R}} g^{(n)}(t) D_n(t| \, b; \tau)\, dt = \int_{\Delta^n} g^{(n)}(\tau \cdot u)\, d\mu_b(u),$$

holds true for $\tau \in \mathbb{R}^{n+1}$ and for all $g \in C^n(\mathbb{R})$, and thus for $g \in \mathscr{S}(\mathbb{R})$.

To define a multivariate analogue of the univariate complex B-splines, we proceed as follows. Let $\lambda \in \mathbb{R}^s \setminus \{0\}$ be a direction, and let $g : \mathbb{R} \to \mathbb{C}$ be a function. The *ridge function* corresponding to g is defined as $g_\lambda : \mathbb{R}^s \to \mathbb{C}$,

$$g_\lambda(x) = g(\langle \lambda, x \rangle) \quad \text{for all } x \in \mathbb{R}^s.$$

We denote the canonical inner product in $\mathbb{R}^s$ by $\langle \bullet, \bullet \rangle$ and the norm induced by it by $\| \bullet \|$.

Definition 3 ([9])**.** Let $\tau = \{\tau^n\}_{n \in \mathbb{N}_0} \in (\mathbb{R}^s)^{\mathbb{N}_0}$ be a sequence of knots in $\mathbb{R}^s$ with the property that

$$\exists\, \varrho \in [0, e) : \limsup_{n \to \infty} \sqrt[n]{\|\tau^n\|} \leq \varrho. \tag{3}$$

The multivariate complex B-spline $\boldsymbol{B}_z(\bullet \mid b, \tau) : \mathbb{R}^s \to \mathbb{C}$ of order z, $\operatorname{Re} z > 1$, with weight vector $b \in \mathbb{C}_+^{\mathbb{N}_0}$ and knot sequence τ is defined by means of the identity

$$\int_{\mathbb{R}^s} g(\langle \lambda, x \rangle) \boldsymbol{B}_z(x \mid b, \tau)\, dx = \int_{\mathbb{R}} g(t) B_z(t \mid b, \lambda\tau)\, dt, \tag{4}$$

where $g \in \mathscr{S}(\mathbb{R})$, and where $\lambda \in \mathbb{R}^s \setminus \{0\}$ such that $\lambda\tau := \{\langle \lambda, \tau^n \rangle\}_{n \in \mathbb{N}_0}$ is separated.

As consequence of the fact that $B_z(\bullet \mid b; \tau) \in L^2(\mathbb{R})$, one obtains from the above definition that $\boldsymbol{B}_z(\bullet \mid b, \tau) \in L^2(\mathbb{R}^s)$ [5]. Moreover, it follows from the Hermite-Genocchi formula for the univariate complex B-splines $B_z(\bullet \,|\, b, \lambda\tau)$ and (4), that $\boldsymbol{B}_z(\, x \,|\, b, \tau) = 0$, when $x \notin [\tau]$, the convex hull of τ.

3. Dirichlet Averages

Let Ω to be a nonempty open convex set in $\mathbb{C}^s$, $s \in \mathbb{N}$, and let $b \in \mathbb{C}_+^{\mathbb{N}_0}$. Let $f \in \mathscr{S}(\Omega) := \mathscr{S}(\Omega, \mathbb{C})$ be a measurable function. For $\tau \in \Omega^{\mathbb{N}_0} \subset (\mathbb{C}^s)^{\mathbb{N}_0}$ and $u \in \triangle^\infty$, define $\tau \cdot u$ to be the bilinear mapping $(\tau, u) \mapsto \sum_{i=1}^\infty u_i \tau^i$. The infinite sum exists if there exists a $\varrho \in [0, e)$ so that

$$\limsup_{n \to \infty} \sqrt[n]{\|\tau^n\|} \leq \varrho. \tag{5}$$

Here, $\| \cdot \|$ now denotes the canonical Euclidean norm on $\mathbb{C}^s$. (See also [6].)

Definition 4. Let $f : \Omega \subset \mathbb{C}^s \to \mathbb{C}$ be a measurable function. The Dirichlet average $F : \mathbb{C}_+^{\mathbb{N}_0} \times \Omega^{\mathbb{N}_0} \to \mathbb{C}$ over $\triangle^\infty$ is defined by

$$F(b; \tau) := \int_{\triangle^\infty} f(\tau \cdot u)\, d\mu_b(u),$$

where $\mu_b = \varprojlim \mu_b^n$ is the projective limit of Dirichlet measures on the n-dimensional standard simplex $\triangle^n$.

We remark that the Dirichlet average is holomorphic in $b \in (\mathbb{C}_+)^{\mathbb{N}_0}$ when $f \in C(\Omega, \mathbb{C})$ for every fixed $\tau \in \Omega^{\mathbb{N}_0}$. (See [2] for the finite-dimensional case and [9] for the infinite-dimensional setting.)

Definition 5. [1] Let $f : \Omega \subset \mathbb{C} \to \mathbb{C}$ be continuous. Let $b \in \mathbb{C}_+^{k+1}$ and $\beta \in \mathbb{C}_+^{\varkappa+1}$. Suppose that for fixed $k, \varkappa \in \mathbb{N}$, $X \in \mathbb{C}^{(k+1)\times(\varkappa+1)}$ and that the convex hull $[X]$ of X is contained in Ω. Then the double Dirichlet average of f is defined by

$$\mathscr{F}(b; X; \beta) := \int_{\triangle^k} \int_{\triangle^\varkappa} f(u \cdot Xv) d\mu_b^k(u) d\nu_\beta^\varkappa(v),$$

where $u \cdot Xv := \sum_{i=0}^k \sum_{j=0}^\varkappa u_i X_{ij} v_j$.

Note that $\mathscr{F}(b; X; \beta)$ is holomorphic on Ω in the elements of b, β, and X.

We again use projective limits to extend the notion of double Dirichlet average to an infinite-dimenional setting. To this end, let $u, v \in \triangle^\infty$ and let $\mu_b = \varprojlim \mu_b^n$ and $\nu_\beta = \varprojlim \nu_\beta^n$ be the projective limits of Dirichlet measures μ_b^n and ν_β^n of the form (1) on the n-dimensional standard simplex, where $b, \beta \in \mathbb{C}_+^{\mathbb{N}_0}$. Now suppose that $X \in \mathbb{C}^{\mathbb{N}_0 \times \mathbb{N}_0}$ is a infinite matrix with the property that $\sum_{i=0}^\infty \sum_{j=0}^\infty |X_{ij}|$ converges. Let

$$u \cdot Xv := \sum_{i=0}^\infty \sum_{j=0}^\infty u_i X_{ij} v_j.$$

Suppose that $\Omega \subset \mathbb{C}$ contains the convex hull $[X]$ of X and that $f : \Omega \to \mathbb{C}$ is continuous. The double Dirichlet average of f over $\triangle^\infty$ is then given by

$$\mathscr{F}(b; X; \beta) := \int_{\triangle^\infty} \int_{\triangle^\infty} f(u \cdot Xv) d\mu_b(u) d\nu_\beta(v). \tag{6}$$

(We use the same symbol for the (double) Dirichlet average over $\triangle^\infty$ and its finite-dimensional projections $\triangle^n$.)

It is easy to show that

$$\mathscr{F}(b; X; \beta) = \int_{\triangle^\infty} F(\beta; uX) d\mu_b(u), \tag{7}$$

where $uX := \{\langle u, X_j \rangle\}_{j \in \mathbb{N}_0}$, with X_j denoting the j-column of X.

We note that $\mathscr{F}(b; X; \beta)$ is holomorphic in the elements of b, β, and X over $\triangle^\infty$.

For $z \in \mathbb{C}_+$, we define

$$\mathscr{F}^{(z)}(b; X; \beta) := \int_{\triangle^\infty} \int_{\triangle^\infty} f^{(z)}(u \cdot Xv) d\mu_b(u) d\nu_\beta(v).$$

(See also [9] for the case of a single Dirichlet average.)

4. Double Dirichlet Averages and Complex B-Splines

Assume now that the matrix X is real-valued and of the form $X_{ij} = 0$, for $i \geq s$ and all $j \in \mathbb{N}_0$, some $s \in \mathbb{N}$. In other words, $X \in \mathbb{R}^{s\times\mathbb{N}_0}$.

Theorem 6. *Suppose that $\beta \in \mathbb{R}^\infty_+$ and that $\operatorname{Re} z > 1$. Let $b := (b_0, b_1, \ldots, b_{s-1}) \in \mathbb{R}^s$ be such that $\sum_{i=0}^{s-1} b_i \notin -\mathbb{N}_0$. Assume that $f \in \mathscr{S}(\mathbb{R}^+)$. Further assume that uX is separated for all $u \in \triangle^{s-1}$. Then*

$$\mathscr{F}^{(z)}(b; X; \beta) = \int_{\mathbb{R}^s} \boldsymbol{B}_z(x \mid \beta, X)\, F^{(z)}(b; x)dx.$$

Proof. We prove the formula first for $b \in \mathbb{R}^s_+$. To this end, we identify $u = (u_0, u_1, \ldots, u_{s-1}, 0, 0, \ldots) \in \triangle^\infty$ with $(u_0, u_1, \ldots, u_{s-1}) \in \triangle^{s-1}$. By the Hermite-Genocchi formula for complex B-splines (see [6] and to some extend [9]), we have that

$$\begin{aligned} F^{(z)}(\beta; uX) &= \int_{\triangle^\infty} f^{(z)}(u' \cdot uX)\, d\mu_\beta(u') \\ &= \int_{\mathbb{R}} f^{(z)}(t) B_z(t \mid \beta, uX)dt \end{aligned}$$

Substituting this expression into (7) and using (4) gives

$$\mathscr{F}^{(z)}(b; X; \beta) = \int_{\triangle^\infty} \int_{\mathbb{R}^s} f^{(z)}(\langle u, x\rangle)\boldsymbol{B}_z(x \mid \beta, uX)\, dx\, d\mu_b(u).$$

Interchanging the order of integration yields the statement for $b \in \mathbb{R}^s_+$. To obtain the general case $b \in \mathbb{R}^s$, we note that by Theorem 6.3-7 in [2], the Dirichlet average F can be holomorphically continued in the b-parameters provided $\sum_{i=0}^{s-1} b_i \notin -\mathbb{N}_0$. □

Remark 7. Theorem 6 extends Theorem 6.1 in [12] to complex B-splines and the $\triangle^\infty$-setting.

5. Moments of Complex B-Splines

Following [2], we define the R-hypergeometric function $R_a(b; \tau) : \mathbb{R}^s_+ \times \Omega^s \to \mathbb{C}$ by

$$R_a(b; \tau) := \int_{\triangle^{s-1}} (\tau \cdot u)^a d\mu_b^{s-1}(u), \tag{8}$$

where $\Omega := H$, H a half-plane in $\mathbb{C} \setminus \{0\}$, if $a \in \mathbb{C} \setminus \mathbb{N}$, and $\Omega := \mathbb{C}$, if $a \in \mathbb{N}$. It can be shown (see [2]) that R_{-a}, $a \in \mathbb{C}_+$, has a holomorphic continuation in τ to $\mathbb{C}_0$, where $\mathbb{C}_0 := \{\zeta \in \mathbb{C} \mid -\pi < \arg\zeta < \pi\}$.
Taking in the definition of the double Dirichlet average (6) for f the real-valued function $t \mapsto t^{-c}$, where $c := \sum_{i=0}^{s-1} b_i$, the resulting double Dirichlet average is denoted by $\mathscr{R}_{-c}(b; X; \beta)$ and generalizes power functions. The corresponding single Dirichlet average $R_{-c}(b; x)$, where $x = (x_0, \ldots, x_{s-1})$, is given by

$$R_{-c}(b; x) = \prod_{i=0}^{s-1} x_i^{-b_i}, \quad x \notin [X]. \tag{9}$$

(See, [2], (6.6-5).)
Now, let $p = (p_0, p_1, \ldots, p_{s-1}) \in \mathbb{R}^s$, $s \in \mathbb{N}$, be a multi-index all of whose components satisfy $p_i < -\frac{1}{2}$. The moment $\mathsf{M}^{(z)}_{|p|}(b; X)$ of order $|p| := \sum_{i=1}^s p_i$ of the complex B-spline $\boldsymbol{B}_z(\bullet \mid \beta, X)$ is defined by

$$\mathsf{M}^{(z)}_{|p|}(b; X) := \int_{\mathbb{R}^s} x^p\, \boldsymbol{B}_z(x \mid \beta, X)\, dx.$$

Note that since $\boldsymbol{B}_z(\bullet \mid \beta, X) \in L^2(\mathbb{R}^s)$ and $\boldsymbol{B}_z(\bullet \mid \beta, X) = 0$, for $x \notin [X]$, an easy application of the Cauchy-Schwartz inequality shows that the above integral exists provided the multi-index p satisfies the aforementioned condition on its components.
Using a result from [8], namely Property 2.5 (b), and requiring that $\operatorname{Re} z < \operatorname{Re} c$, we substitute the function $f := \frac{\Gamma(c-z)}{\Gamma(c)}\,(\bullet)^{-(c-z)}$ into (8) to obtain

$$R^{(z)}_{-(c-z)}(b; x) = R_{-c}(b; x) = \prod_{i=0}^{s-1} x_i^{b_i}.$$

The above considerations together with Theorem 6 immediately yield the next result.

Corollary 8. *Suppose that $\beta \in \mathbb{R}^\infty_+$ and that $\operatorname{Re} z > 1$. Let $b := (b_0, b_1, \ldots, b_{s-1}) \in (-\infty, -\frac{1}{2})^s$ be such that $c := \sum_{i=0}^{s-1} b_i \notin -\mathbb{N}_0$. Moreover, suppose that $\operatorname{Re} z < \operatorname{Re} c$. Then*

$$\mathsf{M}^{(z)}_{-c}(b; X) = \mathscr{R}^{(z)}_{-(c-z)}(b; X; \beta). \tag{10}$$

6. Acknowledgements

This work was partially supported by the grant MEXT-CT-2004-013477, Acronym MAMEBIA, of the European Commission.

References:

[1] B. C. Carlson. Appell functions and multiple averages. *SIAM J. Math. Anal.*, 2(3):420–430, August 1971.

[2] B. C. Carlson. *Special Functions of Applied Mathematics*. Academic Press, New York, 1977.

[3] B. C. Carlson. B-splines, hypergeometric functions, and Dirichlet averages. *J. Approx. Th.*, 67:311–325, 1991.

[4] B. Forster, T. Blu, and M. Unser. Complex B-splines. *Appl. Comp. Harmon. Anal.*, 20:261–282, 2006.

[5] B. Forster and P. Massopust. Multivariate complex B-splines, Dirichlet averages and difference operators. *accepted SAMPTA 2009*, 2009.

[6] B. Forster and P. Massopust. Statistical encounters with complex B-splines. *Constr. Approx.*, 29(3):325–344, 2009.

[7] S. Karlin, C. A. Micchelli, and Y. Rinott. Multivariate splines: A probabilistic perspective. *Journal of Multivariate Analysis*, 20:69–90, 1986.

[8] A. A. Kilbas, H. M. Srivastava, and J. J. Trujillo. *Theory and Applications of Fractional Differential Equations*. Elsevier B. V., Amsterdam, The Netherlands, 2006.

[9] P. Massopust and B. Forster. Multivariate complex B-splines and Dirichlet averages. *to appear in J. Approx. Th.*, 2009.

[10] C. A. Micchelli. A constructive approach to Kergin interpolation in $\mathbb{R}^k$: Multivariate B-splines and Lagrange interpolation. *Rocky Mt. J. Math.*, 10(3):485–497, 1980.

[11] K. S. Miller and B. Ross. *An Introduction to the Fractional Calculus and Fractional Differential Equations*. Wiley, New York, 1993.

[12] E. Neuman and P. J. Van Fleet. Moments of Dirichlet splines and their applications to hypergeometric functions. *J. Comput. and Appl. Math.*, 53:225–241, 1994.

[13] S. G. Samko, A. A. Kilbas, and O. I. Marichev. *Fractional Integrals and Derivatives*. Gordon and Breach Science Publishers, Minsk, Belarus, 1987.

Sampling in cylindrical 2D PET

Yannick Grondin[(1,2)], Laurent Desbat[(1)] and Michel Desvignes[(2)]
(1) TIMC-IMAG, UMR CNRS 5525, UJF-Grenoble 1 (GU) In^{3}S, Faculté de Médecine, 38706 La Tronche France
(2) Grenoble-INP/Phelma/ GIPSA-LAB
961 Rue de la houille blanche BP 46 St Martin d'Heres France
Yannick.Grondin@imag.fr, Laurent.Desbat@imag.fr, michel.desvignes@gipsa-lab.inpg.fr

Abstract:

In this paper, we study 2D cylindrical Positron Emission Tomography (2D PET) sampling. We show that rectangular sampling schemes are more efficient than usual square schemes.

1. PET and sampling

1.1 PET

The aim of Positron Emission Tomography (PET) is to map the internal nuclear activity of a patient from exterior measurement. Usually, the patient received some nuclear substance by inhalation or injection. In PET this substance is tagged with a radioactive isotope, such as Carbon-11, Fluorine-18, Oxygen-15. This substance has also chemical and biological properties that enable to visualize metabolism and functions of patient organs (such as blood flow). This substance, called radiotracer, emits a positron per decay. The positron annihilates with an electron, which results in the emission of two opposite gamma rays detected in a PET system. Thanks to detectors surrounding the patient and a powerful electronic processing, coincident photon pairs can be sorted, meaning that the emission occurred on the line joining both detectors.

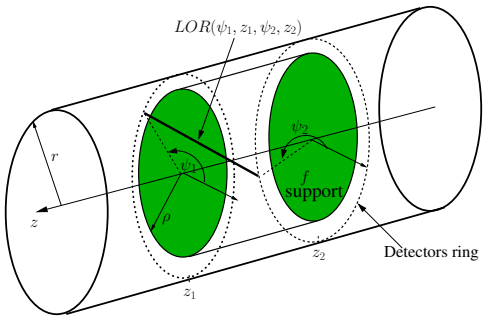


Figure 1: Parametrization of a LOR with the variables $(\psi_1, z_1, \psi_2, z_2)$.

In a cylindrical PET system of radius r, see Fig. 1, the unitary detectors are distributed on a cylinder surrounding the patient (supposed to lie in a cylinder of radius ρ). Each gamma ray detector localization can be parametrized by cylindrical coordinates (ψ, z). When the coincidence on two detectors (ψ_1, z_1) and (ψ_2, z_2) is detected, one knows that some activity occurs on the line joining the detectors (ψ_1, z_1) and (ψ_2, z_2). This line is called a LOR (Line Of Response).

In 2D mode, lead rings called septa, see Fig. 2, are used to restrict detected LORs to be essentially perpendicular to the PET cylinder axis. In this case, LORs have only three parameters (ψ_1, ψ_2, z), see Fig. 3. LORs with a small oblicity (crossed LORs) are usually approximated to LORs perpendicular to the axis, between two true detectors rings, creating a virtual detection ring, allowing to improve the sampling rate along the axis direction, see Fig. 2.

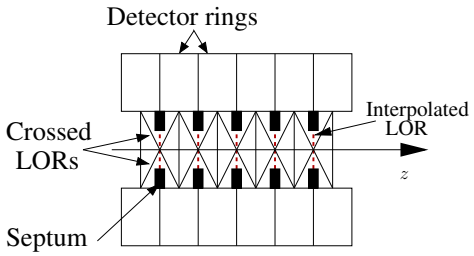


Figure 2: Crossed LORs interpolated to improve axial sampling .

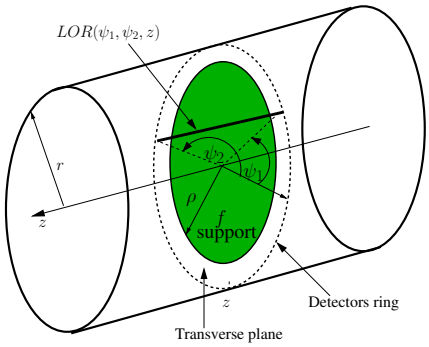


Figure 3: Parametrization of a LOR with the variables (ψ_1, ψ_2, z).

In 2D PET, after the attenuation correction [5] the measure can be modeled by $g : [0, 2\pi] \times [0, 2\pi] \times \mathbb{R} \to \mathbb{R}$, with

$$g(\psi_1, \psi_2, z) = \int_{\mathbb{R}} f(u(\psi_1, z) + t\theta(\psi_1, \psi_2))\, dt$$

with $u(\psi_1, z) = (r\cos\psi_1, r\sin\psi_1, z)^t$ and $\theta(\psi_1, \psi_2) =$

$\frac{1}{2\left|\sin\left(\frac{\psi_1-\psi_2}{2}\right)\right|}(\cos\psi_2-\cos\psi_1,\sin\psi_2-\sin\psi_1,0)^t$. Obviously g satisfies the symmetry relation

$$g(\psi_1,\psi_2,z)=g(\psi_2,\psi_1,z). \tag{1}$$

1.2 Sampling

We want to sample a function g being 2π-periodic in its two first variables and in $\mathbb{R}$ in its third variable. This is a particular case of the general framework of sampling of function on groups, see for example [2, 3]. In this case, the Fourier transform of $g\in C_0^\infty([0;2\pi[\times[0;2\pi[\times\mathbb{R})$ can be defined by:

$$\hat{g}(\xi)=\frac{1}{(2\pi)^2\sqrt{2\pi}}\int_{[0;2\pi[}\int_{[0;2\pi[}\int_{\mathbb{R}}g(x)e^{-ix\cdot\xi}dx,$$

where $x=(\psi_1,\psi_2,z)^t\in[0;2\pi[\times[0;2\pi[\times\mathbb{R}$, $\xi=(p_1,p_2,\zeta)^t\in\mathbb{Z}\times\mathbb{Z}\times\mathbb{R}$ and $\xi\cdot x=p_1\psi_1+p_2\psi_2+\zeta z$. The inverse Fourier transform defined for G a function on $\mathbb{Z}\times\mathbb{Z}\times\mathbb{R}$ is given by

$$\begin{aligned}\check{G}(x)&=\frac{1}{\sqrt{2\pi}}\int_{\mathbb{Z}\times\mathbb{Z}\times\mathbb{R}}G(\xi)e^{ix\cdot\xi}\\&=\frac{1}{\sqrt{2\pi}}\sum_{p_1\in\mathbb{Z}}\sum_{p_2\in\mathbb{Z}}\int_{\zeta\in\mathbb{R}}G(p_1,p_2,\zeta)e^{i(p_1\psi_1+p_2\psi_2+\zeta z)}d\zeta.\end{aligned}$$

Let $\mathbf{K}\subset\mathbb{Z}\times\mathbb{Z}\times\mathbb{R}$, the non-overlapping Shannon condition associated to $\mathbf{K}$ for the sampling lattice $L_W=W\mathbb{Z}^3\cap([0;2\pi[\times[0;2\pi[\times\mathbb{R})$ generated by the non singular 3×3 matrix W is that *the sets* $\mathbf{K}+2\pi W^{-t}l, l\in\mathbb{Z}^3$ *are disjoint sets in* $\mathbb{Z}\times\mathbb{Z}\times\mathbb{R}$. The Petersen-Middleton theorem [6, 3] yields the Fourier interpolation formula

$$(S_Wg)(x)=\frac{1}{\sqrt{2\pi}}|\det W|\sum_{y\in L_W}g(y)\check{\chi}_{\mathbf{K}}(x-y),$$

where $\chi_{\mathbf{K}}$ is the indicator function of the set $\mathbf{K}$. The interpolation error is given by

$$||S_Wg-g||_\infty\le\frac{2}{\sqrt{2\pi}}\int_{\xi\notin\mathbf{K}}|\hat{g}(\xi)|d\xi.$$

Thus if $\mathbf{K}$ is the essential support of $\hat{g}$, i.e., $\int_{\xi\notin\mathbf{K}}|\hat{g}(\xi)|d\xi$ can be negligible, then the interpolation error is low. The geometry of the set $\mathbf{K}$ can be exploited for the design of efficient sampling schemes, i.e., the choice of W satisfying the Shannon condition with $|\det W|$ maximal in order to minimize the number of sampling points.

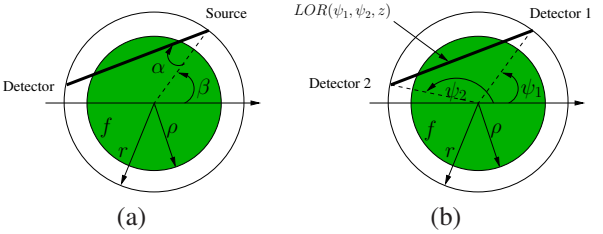


Figure 4: Fan beam (a) and natural PET (b) parametrization in a transverse plane .

2. 3D Sampling in cylindrical PET 2D mode

In [1] we have established the sampling conditions of the 3D Fan-Beam X-ray Transform (3DFBXRT):

$$\mathcal{D}_{e_3{}^\perp}f(\beta,\alpha,z)=\int_{L_{\beta,\alpha,t}}f(u)du,$$

where $u\in\mathbb{R}^3$, $L_{\beta,\alpha,z}$ is the line in the plane perpendicular to e_3 at abscissa z ($z\in\mathbb{R}$), joining the source at $r(\cos\beta,\sin\beta,0)^t+ze_3$, $\beta\in[0,2\pi[$ and the detector at angular position $\alpha\in[-\pi/2,\pi/2[$, see Fig. 4. This geometry appears in X-ray CT scanner when considering the reconstruction of many 2D slices. Cylindrical PET in 2D mode can be linked with the 3DFBXRT in the following way:

$$g(x)=D_{3D}f(\mathbf{A}(x-e_\pi)) \tag{2}$$

where $x=(\psi_1,\,\psi_2,\,z)^t$, $e_\pi=(0,\,\pi,\,0)^t$, and

$$\mathbf{A}=\begin{pmatrix}1&0&0\\-\frac{1}{2}&\frac{1}{2}&0\\0&0&1\end{pmatrix}$$

see Fig 4.

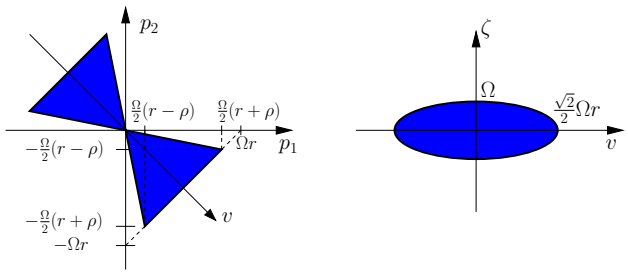


Figure 5: K_g: essential support of $\hat{g}$ for $\eta=\rho/r=2/3$, slices in the planes (p_1,p_2) (left) and (v,ζ) (right). The 3D set K_g is just at the intersection of two cylinders of respective basis the slices in the (p_1,p_2) and (v,ζ) and respective axis ζ and the direction perpendicular to (v,ζ) .

This link allows to easily estimate the essential support of $\widehat{g}:\mathbb{Z}\times\mathbb{Z}\times\mathbb{R}\to\mathbb{R}$. Indeed,

$$\begin{aligned}\widehat{g}(\xi)&=\int_{[0;2\pi[}\int_{[0;2\pi[}\int_{\mathbb{R}}g(x)e^{-ix\cdot\xi}dx\\&=\int_{[0;2\pi[}\int_{[0;2\pi[}\int_{\mathbb{R}}D_{3D}f(\mathbf{A}(x-e_\pi))e^{-ix\cdot\xi}dx\\&=\int_{[0;2\pi[}\int_{[0;2\pi[}\int_{\mathbb{R}}D_{3D}f(\mathbf{A}x)e^{-ix\cdot\xi+ip_2\pi}dx\\&=\frac{(-1)^{p_2}}{|\det\mathbf{A}|}\int_{[0;2\pi[}\int_{[0;2\pi[}\int_{\mathbb{R}}D_{3D}f(x)e^{-i(\mathbf{A^{-1}}x)\cdot\xi}dx\\&=\frac{(-1)^{p_2}}{|\det\mathbf{A}|}\int_{[0;2\pi[}\int_{[0;2\pi[}\int_{\mathbb{R}}D_{3D}f(x)e^{-ix\cdot\left(\mathbf{A^{-t}}\xi\right)}dx\\&=\frac{(-1)^{p_2}}{|\det\mathbf{A}|}\widehat{D_{3D}f}(\mathbf{A^{-t}}(\xi))\end{aligned}$$

From this link we see that the essential support of $\widehat{g}$ is simply a linear transformation of the essential support of

$\widehat{D_{3D}f}$. From [1] it can be easily shown that K_g, the essential support of $\hat{g}(p_1, p_2, \zeta)$ when the emission function f is supposed the be essentially Ω band limited, is given by

$$K_g = \{(p_1, p_2, \zeta) \in \mathbb{Z} \times \mathbb{Z} \times \mathbb{R}, \\ |p_1 - p_2|^2 + r^2\zeta^2 < \Omega^2 r^2; r|p_1 + p_2| < \rho|p_1 - p_2|\}$$

see Fig. 5 for a representation.
The angles ψ_1 and ψ_2 parametrize the same detector ring, thus their sampling must be identical. We consider here only standard sampling, i.e. equidistant sampling along each direction. The most efficient diagonal matrix satisfying the non overlapping Shannon conditions, see Fig. 6, is given by:

$$2\pi \mathbf{W_S^{-t}} = \Omega \begin{pmatrix} r & 0 & 0 \\ 0 & r & 0 \\ 0 & 0 & 2 \end{pmatrix}, \mathbf{W_S} = \frac{2\pi}{r\Omega} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{r}{2} \end{pmatrix} \quad (3)$$

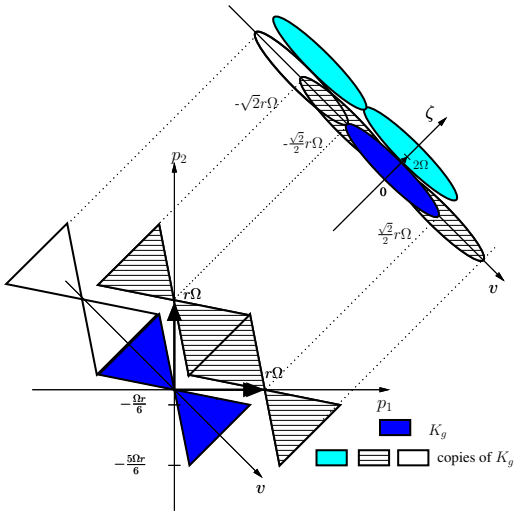


Figure 6: Non overlapping conditions for the rectangular sampling scheme .

Thus we see that the most efficient sampling distances are $\Delta\psi_1 = 2\pi/r\Omega (= \Delta\psi_2)$ and $\Delta z = \pi/\Omega$. $l_z = \Delta z$ would thus be the detector axial length. If we approximate the detector tangential length by $l_t = r\Delta\psi_1$, we see that the most efficient relation is $l_z = l_t/2$, thus the most efficient detectors from the sampling point of view are rectangular detectors. The empirical ring oversampling by rebinning the crossed LORs as in Fig. 2 yields exactly the factor 2 of oversampling in the direction z needed for efficient sampling. This is a theoretical justification of this widely used heuristic rebinning method.

3. Numerical experiments

3.1 Essential support

We have computed from numerical phantom the essential support of $|\hat{g}(p_1, p_2, \zeta)|$ see Fig. 7. In (a) and (c) the simulation is based on simple line integrals of a phantom f built with 3 concentric weighted ball indicator functions: $f = \chi_{B(c,0.03)} + \chi_{B(c,0.05)} + \chi_{B(c,0.07)}$ where $\chi_{B(c,r)}$ is the indicator function of the ball of radius r centered on $c = (0.9, 0, 0)$. The data are simulated for a PET of radius 1.5 with 32 rings and 300 detectors on each ring. (b) and (d) are based on a Monte Carlo (MC) simulation computed with GATE [4]. The phantom f is built with 5 concentric weighted ball sources (of radius r expressed in mm): $f = a(\chi_{B(c,9)} + \chi_{B(c,10)} + \chi_{B(c,11)} + \chi_{B(c,12)} + \chi_{B(c,13)})$, where the center $c = (130, 0, 0)$ mm and the activity $a = 10^6$ becquerel. The data are simulated for a PET of radius 402 mm with 32 rings and 576 detectors on each ring, imitating the ECAT EXACT HR$^+$ scanner of CTI/Siemens. We see that the simulation data are in good agreement with the theoretical results.

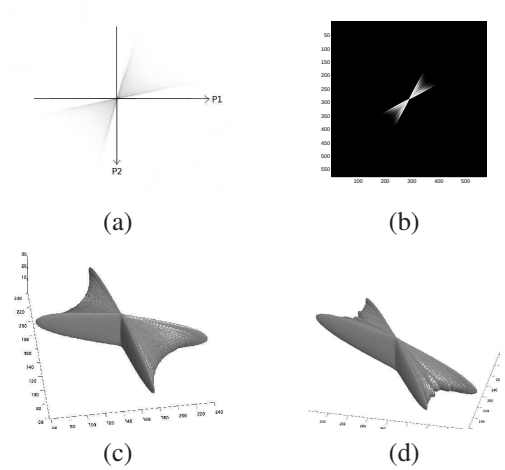


Figure 7: In (a) and (c) the emission function f is the sum of 3 concentric indicator functions. In (b) and (c) the data are obtained by a MC simulation of 5 concentric spherical sources. (a) and (b) slice $\zeta = 0$ of $|\hat{g}(p_1, p_2, \zeta)|$; (c) and (d) 3D visualization of the isosurface at 1% of maximum of $|\hat{g}(p_1, p_2, \zeta)|$ ($|\hat{g}(p_1, p_2, \zeta)|$ is essentially negligible outside of this surface) .

3.2 Reconstruction resolution

In Fig. 8, Fig. 9 and 10, we present the reconstruction of the clock phantom, see [8], from simple line integrals. The simulated cylindrical PET is of radius $r = 1.5$, the reconstruction region is of radius $\rho = 1$. We consider two sampling schemes with essentially the same number of data. The square scheme is based on square detectors, with $l_t = l_z = 0.049$. The number of ring is 20. The number of detectors on a ring is 190. The rectangular scheme is based on rectangular detectors, with $l_t = 2l_z = 0.062$. The number of ring is 32. The number of detectors on a ring is 150. We see in these numerical experiments that the rectangular scheme yields better reconstructions than the square scheme.

4. Conclusion

We have shown the efficiency of the rectangular sampling scheme over the square scheme in 2D mode cylindrical PET. Sampling conditions in fully 3D PET as initiated in [7] are now being investigated.

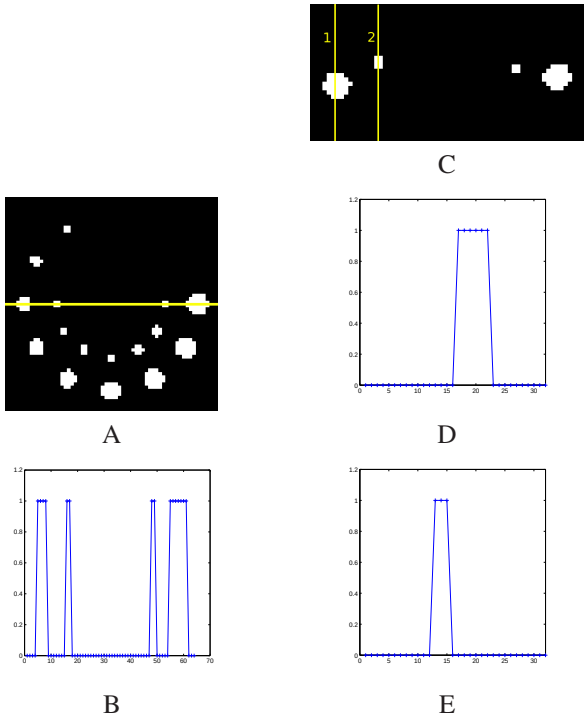


Figure 8: A = Original image: transverse view ; B = Image profile ; C = Original image: axial view ; D = Image profile 1 ; E = Image profile 2 .

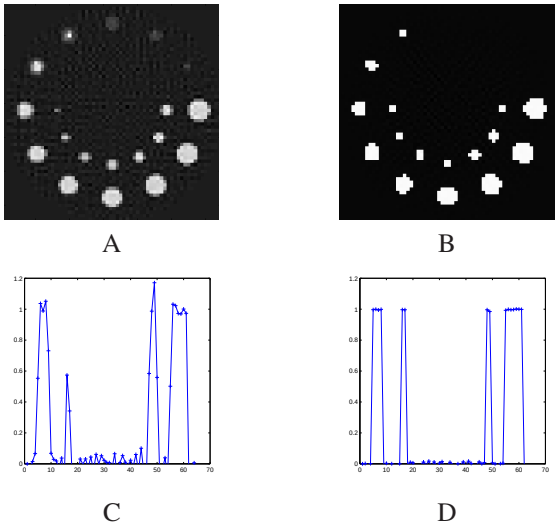


Figure 9: A = Square scheme image: transverse view ; B = Rectangular scheme image: transverse view ; C = Square scheme image profile ; D = Rectangular scheme image profile .

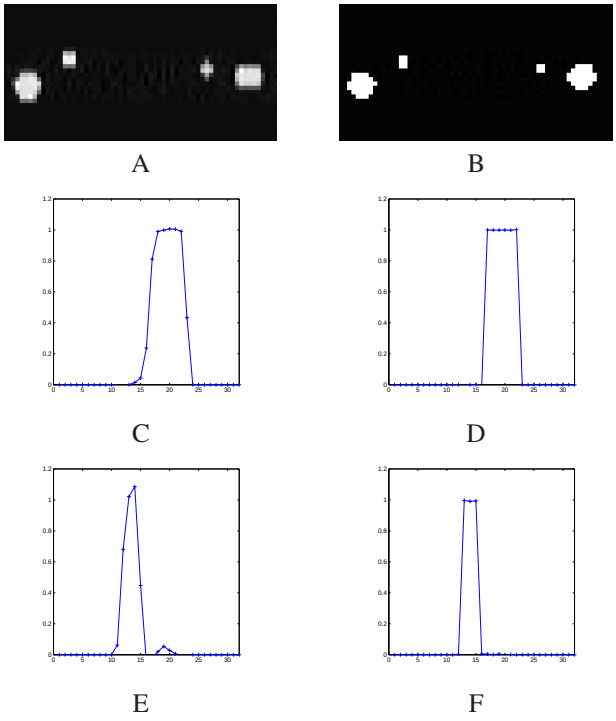


Figure 10: A = Square scheme image: axial view ; B = Rectangular scheme image: axial view ; C = Square scheme image profile 1 ; D = Rectangular scheme image profile 1 ; E = Square scheme image profile 2 ; F = Rectangular scheme image profile 2 .

References:

[1] L. Desbat, S. Roux, P. Grangeat, and A. Koenig. Sampling conditions of 3D Parallel and Fan-Beam X-ray CT with application to helical tomography. *Phys. Med. Biol.*, 49(11):2377–2390, 2004.

[2] A. Faridani. An application of a multidimensional sampling theorem to computed tomography. In *AMS-IMS-SIAM Conference on Integral Geometry and Tomography*, volume 113, pages 65–80. Contemporary Mathematics, 1990.

[3] A. Faridani. A generalized sampling theorem for locally compact abelian groups. *Math. Comp.*, 63(207):307–327, 1994.

[4] S. Jan and coll. Gate: a simulation toolkit for pet and spect. *Phys. Med. Biol*, 49:4543–4561, 2004.

[5] F. Natterer. *The Mathematics of Computerized Tomography*. Wiley, 1986.

[6] D.P. Petersen and D. Middleton. Sampling and reconstruction of wavenumber-limited functions in N-dimensional euclidean space. *Inf. Control*, 5:279–323, 1962.

[7] T. Rodet, J. Nuyts, M. Defrise, and C. Michel. A study of data sampling in pet using planar detectors. In *IEEE Nuclear Science Symp. Conf. Rec*, 2003.

[8] Henrik Turbell. *Cone-Beam Reconstruction Using Filtered Backprojection*. PhD thesis, Linkping University, 2001.

Significant Reduction of Gibbs' Overshoot with Generalized Sampling Method

Yufang Hao[(1)], Achim Kempf [(1),(2)]

(1) Department of Applied Mathematics, University of Waterloo, Waterloo, ON, N2L 3G1, Canada
(2) Department of Physics, University of Queensland, St Lucia 4072, QLD, Australia
yhao@math.uwaterloo.ca

Abstract:

As is well-known, the use of Shannon sampling to interpolate functions with discontinuous jump points leads to the Gibbs' overshoot. In image processing, it can lead to the problem of artifacts close to edges, known as Gibbs ringring. Its amplitude cannot be reduced by increasing the sample density. Here we consider a generalized Shannon sampling method which allows the use of time-varying sample densities so that samples can be taken at a varying rate adapted to the behavior of the function. Using this generalized sampling method to approximate a periodic step function, we observe a strong reduction of Gibbs' overshoot. In a concrete example, the amplitude of the Gibbs' overshoot is reduced by about 70%.

1. Introduction

The Shannon sampling theorem [6] provides the link between continuous and discrete representations of information and has numerous practical uses in communication engineering and signal processing. For a review on Shannon sampling, see [7, 10, 1]. In addition, the Shannon sampling theorem has been used to interpolate samples to approximate a given function.

In the use of Shannon sampling to approximate functions with discontinuous jump points, the well-known Gibbs' overshoot [2, 3] has remained a persistent problem, leading to, e.g., Gibbs ringing in image compression [5]. The clearest example for the Gibb's phenomenon is the periodic step function $H(t)$, see Figure 1, where $H(t) = 1$ on $(0, \frac{1}{2})$, $H(t) = -1$ on $(\frac{1}{2}, 1)$, $H(t) = 0$ at $t = 0$, $\frac{1}{2}$, 1, and $H(t)$ has a period $T = 1$.

In Figure 1, $H(t)$ is approximated using Shannon's shifted sinc reconstruction kernel with $N = 24$ sampling points on one periodic interval $[0, 1)$. Samples are denoted by $\times$ in the plot, and the solid line at the top indicates the maximum value of the approximating function, which is 1.0640. Within an error of 0.003, the 6.40% overshoot beyond the maximum amplitude 1 of the step function $H(t)$ can not be further reduced even if we increase the sampling density.

However, using the generalized sampling method [4, 8, 9], which allows the reconstruction of a function on a set of non-equidistant sampling points, chosen adaptively according to the behavior of the function, we show that the Gibbs overshoot can be strongly reduced. For an example, see Figure 2.

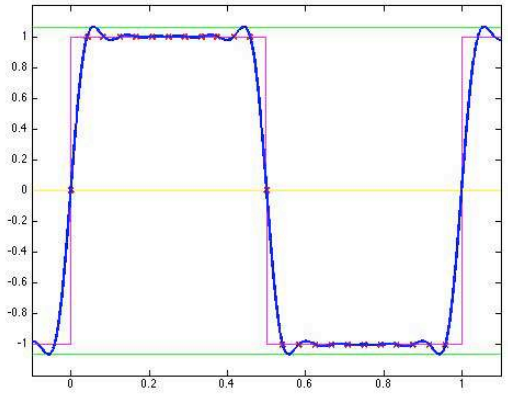


Figure 1: Approximation of the step function by Shannon sampling.

In Figure 2, we use the same number of points $N = 24$ in one period as in the case of Shannon in Figure 1, but we choose the sampling points to match the behaviour of the step function. Intuitively, the jump in the step function contains high frequencies. Thus more samples are taken near the jump points $t = 0$, $\frac{1}{2}$, and 1. In this example, the maximum value of the approximation is reduced to 1.0074 with an error of 0.0003. This is roughly a 70% reduction of Gibbs' overshoot without increasing the number of samples, but only varying the local sample density.

Figure 3 is a zoom-in of Figure 2 near the jump point. The dashed line on the top indicates the maximum values of the approximating function using the generalized sampling, while the solid line indicates the overshoot in the case of Shannon.

2. Generalized Shannon Sampling Method

The generalized Shannon sampling theory considered here was not specifically developed for the application of reducing Gibbs' phenomena. It was originally motivated by some fundamental physics problem in quantum gravity [4] and was introduced to engineering for spaces of functions with a new notion of time-varying Nyquist rate [8, 9]. The starting observation is that each set of Nyquist sampling points in Shannon sampling turns to be the eigen-

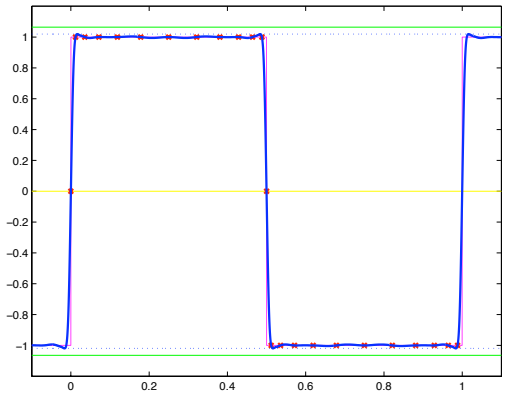


Figure 2: Approximating the step function by the generalized sampling method with non-equidistant sampling points..

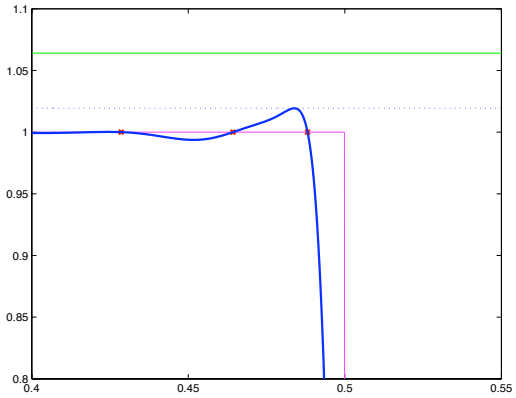


Figure 3: This is a zoom-in of Figure 2 near the jump point..

values of one of the self-adjoint extensions of a particular simply symmetric multiplication operator T with deficiency indices $(1,1)$, and the shifted sinc kernels are the corresponding eigenfunctions. The Shannon sampling theorem is the special case when the self-adjoint extensions of T have equidistant eigenvalues. By considering a generic such symmetric operator T, one obtains a generalized sampling method. We can not cover the mathematical derivations of the new generalized sampling method here, but we will review the key features of the generalization along with a comparison to the Shannon sampling theorem.

The Shannon sampling theorem states that if a function $\phi(t)$ is in the space of Ω-bandlimited functions, i.e., $\phi(t)$ has a frequency upper bound Ω, then $\phi(t)$ can be perfectly reconstructed from its sample values $\{\phi(t_n)\}_n$ taken on a set of sampling points $\{t_n\}_n$ with an equidistant spacing $t_{n+1} - t_n = 1/(2\Omega)$ via:

$$\phi(t) = \sum_{n=-\infty}^{\infty} G\big(t, t_n\big)\phi(t_n) \qquad (1)$$

The function $G(t, t_n)$ is the so-called reconstruction kernel, which is the shifted sinc function $\mathrm{sinc}\big(2\Omega(t - t_n)\big)$. The frequency upper bound Ω is called the bandwith, and the sampling rate $1/(2\Omega)$ is the Nyquist sampling rate.

2.1 One-Parameter Family of Sampling Lattices

We will call a set of Nyquist sampling points $\{t_n\}_n$ a sampling lattice. The Shannon sampling theorem only specifies the constant spacing between adjacent points in one lattice, but it does not specify an initial sampling point. Therefore, we can parameterize all possible sampling lattices as:

$$t_n(\theta) = \frac{n+\theta}{2\Omega}, \quad 0 \le \theta < 1 \qquad (2)$$

Hence the Shannon sampling method possesses a natural one-parameter family of sampling lattices, and any function in the function space can be perfectly reconstructed from its values on any fixed lattice via Eq. (1).

The generalized sampling method also possesses an analogous one-parameter family of sampling lattices, but the points in each lattice are generally non-equidistant now. To distinguish from the case of Shannon, we use a different parameter α in $\{t_n(\alpha)\}_n$, $0 \le \alpha < 1$, and assume that $\{t_n(\alpha)\}_n$ are differentiable with respect to the parameter α:

$$t_n'(\alpha) = \frac{dt_n(\alpha)}{d\alpha}$$

Shannon's family of sampling lattices $\{t_n(\theta)\}_n$ can be generated by a single number, namely, the constant bandwidth Ω. It is so simple because the function space in the case of Shannon has a constant bandwidth Ω. However, in the generalization, since we have a time-varying 'bandwidth', in the sense of Nyquist lattices with non-equidistant points, more specification is required. The entire family of sampling lattices is now generated from the knowledge of a given lattice, say $\{t_n(0)\}$, and a set of corresponding derivatives $\{t_n'(0)\}_n$ by solving for $t = t_n(\alpha)$ in:

$$\sum_m \frac{t_m'(0)}{t - t_m(0)} = \pi\, \cot(\pi\alpha) \qquad (3)$$

This equation implies that one sampling lattice and the correponding derivatives are enough to determine the entire family of sampling lattices, and hence the reconstruction kernel and the function space. This is important for practical purposes, because one usually takes samples of a given signal on only one lattice.

The family of sampling lattices $\{t_n(\alpha)\}_n$ in the generalization shares many important properties of the uniform lattices $\{t_n(\theta)\}$ of Shannon: as the parameter α (or β in the case of Shannon) increases from 0 to 1, the sampling lattices specified by the parameter move to the right on the real line simultaneously and continuously with the following continuity condition:

$$t_n(1) := \lim_{\alpha \to 1^-} t_n(\alpha) = t_{n+1}(0), \;\; t_n'(1) = t_{n+1}'(0) \qquad (4)$$

Hence, together, these sampling points in all lattices again cover the real line exactly once. Namely, for any $t \in \mathbb{R}$, there exists an unique integer n and an unique α in $[0, 1)$ such that $t = t_n(\alpha)$.

2.2 The Generalized Reconstruction Kernel

From the theory of self-adjoint extensions, if on each fixed but arbitrary lattice $\{t_n(\alpha)\}$, α fixed, we let $t_n = t_n(\alpha)$,

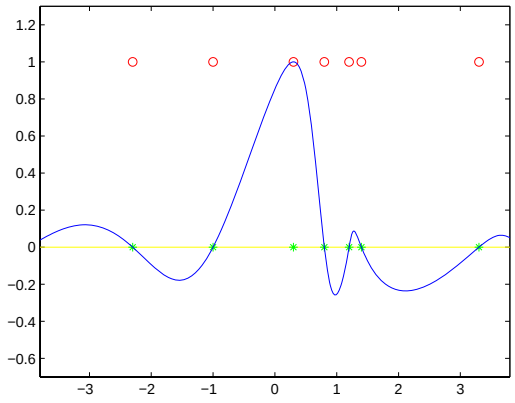


Figure 4: An example of generalized sinc function (or reconstruction kernel) on an arbitrary non-equidistant sampling lattice. The stars on the real line indicate the points in an arbitrary non-equidistant sampling lattice, and the circles denote the same set of points with an amplitude 1.

$t'_n = t'_n(\alpha)$, then the reconstruction kernel in the generalized sampling theorem reads:

$$G(t,t_n) = (-1)^{z(t,\,t_n)}\,\frac{\sqrt{t'_n}}{|\,t-t_n|}\left(\sum_m \frac{t'_m}{(t-t_m)^2}\right)^{-1/2} \tag{5}$$

where $z(t,t_n)$ is the number of the sampling points $\{t_m\}_m$ between t and t_n exclusively.
As functions in t, for each fixed α, the set of functions

$$\left\{g_n^{(\alpha)}(t) = G(t,t_n(\alpha))\right\}_n \tag{6}$$

forms a basis of the function space. Thus, indeed, in the generalized sampling theorem, every function in the function space specified by the family of sampling lattices can be expanded using these basis functions.
These continuous functions in Eq. (6) have analogous properties to the shifted sinc function of Shannon: they interpolate all the points in the lattice specified by α

$$g_n^{(\alpha)}(t_m(\alpha)) = G\big(t_m(\alpha),t_n(\alpha)\big) = \delta_{mn}$$

and their maximum values are all 1 at the sampling points about which they are 'centered'. This is important for the stability of reconstruction. We will refer to these basis functions as generalized sinc functions. See Figure 4 for an example.
It is important to recall that each set of basis functions $\left\{g_n^{(\alpha)}(t)\right\}_n$ specified by α spans the same function space. This property is remarkable since as in Figure 4, the shape of the generalized sinc functions is quite non-trivial.
To recover the Shannon sampling theorem as a special case, we choose any uniform sampling lattice $\{t_n\}_n$ with $t_{n+1}-t_n = \frac{1}{2\Omega}$ for all n, together with constant derivatives $t'_n = C$. Then the reconstruction kernel in (5) simplifies to the sinc kernel $\mathrm{sinc}\big(2\Omega(t-t_n)\big)$, by using the following trignometric identity:

$$\frac{\pi^2}{\sin^2(\pi z)} = \sum_{k=-\infty}^{+\infty}\frac{1}{(z-k)^2} \tag{7}$$

2.3 Interpolation Strategy

To approximate a given function, depending on the behavior of the function, one must select a sampling lattice for interpolation. Arising from the theory of self-adjoint extensions, the chosen lattice $\{t_n\}_n$ must have a minimum and maximum spacing, namely, there must exist positive real numbers $\delta_{\min}$ and $\Delta_{\max}$ such that:

$$0 < \delta_{\min} \le \Delta t_n = t_{n+1}-t_n \le \Delta_{\max} \quad \text{for all } n \tag{8}$$

From Section 2.1 Eq. (3), we know that one also needs a set of corresponding derivatives to apply the generalized sampling method. So the question is, for a given lattice $\{t_n\}_n$, what is a suitable choice of the set of corresponding derivatives $\{t'_n\}_n$?
To this end, we notice that the derivative $t'_n(\alpha)$ is the velocity with which the sampling points $t_n(\alpha)$ are moving to the right along the real line for increasing α at $t = t_n(\alpha)$. Hence, a good candidate for t'_n is the distance travelled in one period of α, which is the spacing between two adjacent points $\Delta t_n = t_{n+1}-t_n$. For symmetry, we set t'_n to be the average distance between t_n to its previous and successive points:

$$t'_n = \frac{1}{2}\Big(\Delta t_n + \Delta t_{n-1}\Big) = \frac{1}{2}\Big(t_{n+1}-t_{n-1}\Big) \tag{9}$$

Here a constant prefactor for the derivatives on a fixed lattice does not matter because the reconstruction kernel is independent of a scalar multiplication of the derivatives: in (5), the prefactor in $\sqrt{t'_n}$-term will cancel out the one in t'_m on the numerator inside the series.
With this set of initial data $\{t_n\}_n$ and $\{t'_n\}_n$, we have an explicit expression of the reconstruction kernel (5). Hence we can construct the interpolating function $\phi(t)$ through all the sample points $\big\{\big(t_n,\phi(t_n)\big)\big\}_n$ using the reconstruction formula (1).

3. Reduction of Gibbs' Overshoot

3.1 Reconstruction of Periodic Functions

The clearest example to demonstrate the reduction of Gibb's overshoot using the generalized sampling method is the periodic step function $H(t)$. One of the reasons for choosing a periodic function is that the infinite summations in the both reconstruction kernel (5) and the reconstruction formula (1) will simplify to a finite sum. Hence, we eliminate the truncation error in the summation.
To this end, assume that the function $\phi(t)$ has a period of T, and we take N sampling points on one period $[0,T)$, which are denoted by $\{\tau_1,\tau_2,\ldots,\tau_N\} \subseteq [0,T)$. Hence, all the sampling points are

$$t_{nN+k} = nT+\tau_k, \quad 1 \le k \le N, n \in \mathbb{N} \tag{10}$$

and from the periodictiy, we have

$$t'_{nN+k} = t'_k, \quad \phi(t_{nN+k}) = \phi(t_k) \tag{11}$$

After a lengthy calculation, the reconstruction kernel (5)

on this periodic lattice now reads:

$$G(t,t_{nN+K}) = \frac{(-1)^{z(t,t_{nN+l})}\sqrt{t'_k}}{|t-t_{nN+K}|} \frac{\pi}{T}\Big(\sum_{l=1}^{N} t'_l \sin^{-2}\big(\frac{\pi}{T}(t-\tau_l)\big)\Big)^{-1/2} \quad (12)$$

and the reconstruction formula (1) of the T-periodic function $\phi(t)$ reads:

$$\phi(t) = \sum_{k=1}^{N} (-1)^{z(t,t_{nN+k})}\sqrt{t'_k}\,\cot\big(\frac{\pi}{T}(t-\tau_k)\big) \Big(\sum_{l=1}^{N} t'_l \sin^{-2}\big(\frac{\pi}{T}(t-\tau_l)\big)\Big)^{-1/2}\phi(t_k) \quad (13)$$

As discussed in Section 2.3, for using the formulae (12) and (13) to approximate a periodic step function $H(t)$, the only task now left is to find a sampling lattice adapted to the behavior of $H(t)$. With a periodic lattice (10), we only need to pick up a finite number N of them on $[0,T)$.

3.2 Approximating a Periodic Step Function

Before we discuss how to determine a set of non-equidistant sampling points, let us first consider why the uniform lattices of Shannon do not work very well. Intuitively, because of the sudden change in the amplitude of a step function $H(t)$ at its jump points $t = 0$, $\frac{1}{2}$ and 1, the function can be considered to suddenly oscillate at an "infinite" frequency in a sufficiently small neighborhoods at the jump points, namely to have an 'infinite' bandwidth at $t = 0$, $\frac{1}{2}$ and 1. Recall that the constant Nyquist spacing $1/(2\Omega)$ in the case of Shannon is inversely proportional to the bandwidth Ω. A uniform lattice implies uniform bandwidth. Intuitively, the uniform lattice in the case of Shannon is therefore not matched with the increase of bandwidth in the small neighborhoods of jump points.

We therefore choose N sampling points with non-equidistant spacings so that the smallest spacing (the highest bandwidth) occurs near the jump points at $t = 0$, $\frac{1}{2}$, 1, and the spacing gradually increases away from the jump points (the bandwidth decreases). We used the easiest such increasing change in spacing, which is linear.

Due to the symmetry of the jump points at $t = 0$, $\frac{1}{2}$, 1, we divide one period $[0,1)$ into four equal subintervals with length $\frac{1}{4}$. On the first subinterval, $[0,\frac{1}{4})$, we choose K points so that their adjacent spacing is linearly increasing. Let δ be the linear increment in spacing, then

$$\tau_1 = 0, \tau_2 = \delta, \tau_3 = 3\delta, \ldots \quad \tau_K = \frac{1}{2}K(K-1)\delta \quad (14)$$

The $(K+1)^{\text{st}}$ point is $\frac{1}{4}$. The sampling points on $(\frac{1}{4},\frac{1}{2}]$ are a mirror image of the points on $[0,\frac{1}{4})$ with respect to $t = \frac{1}{4}$, and the points on $[\frac{1}{2},1)$ repeat the ones on $[0,\frac{1}{2})$. Therefore, we have in total $N = 4K$ points on $[0,1)$.

The approximation in Figure 2 is obtained in this way with $K = 6$. Hence it has the same total number of sampling points ($N = 24$) on $[0,1)$ as in Figure 1. Its maximum amplitude is 1.0193, which is a significant reduction compared to the maximum amplitude 1.0640 in Gibbs' overshoot (Figure 1).

4. Outlook

The question arises how far one can ultimately reduce the Gibb's overshoot? Is the linear change in sampling spacing, as in Eq. (14), the optimal lattice spacing to match the behavior of a step function? This question will be addressed in a longer following-up paper, in which we will pursue an analytical optimization of the Gibbs' overshoot reduction.

To this end, the fact that the closed form of the reconstruction kernel (12) is available in the case of periodic functions has an important advantage: it in effect reduces infinitely many points to a set of finitely many points. We can then analytically study the behavior of the constructed approximating functions. Eventually, we hope such an analytical study can lead us to the ultimately goal, which is to provide solution to design optimally adapted lattices for arbitrary given functions.

5. Acknowledgment

This work has been supported by NSERC's Discovery, Canada Research Chairs and CGS D2 programs. A.K. and Y.H. gratefully acknowledge the kind hospitality at the University of Queensland, where A.K. is currently on sabbatical.

References:

[1] J.J. Benedetto. *Modern Sampling Theory*. Birkhauser, Boston, 2001.

[2] J.W. Gibbs. Fourier series. *Nature*, 59:606, 1899.

[3] A.J. Jerri. *The Gibbs Phenomenon in Fourier Analysis, Splines and Wavelet Approximations*. Springer, 1998.

[4] A. Kempf. On fields with finite information density. *Phys. Rev. D*, 69(124014), 2004.

[5] A. Gelb R. Archilbald. A method to reduce the gibbs ringing artifact in mri scans while keeping tissue boundary integrity. *IEEE Trans. on Medical Imaging*, 21(4):305–319, Apr. 2002.

[6] C.E. Shannon. Communication in the presence of noise. *Proc. IRE*, 37:10–21, Jan. 1949.

[7] M. Unser. Sampling - 50 years after shannon. *Proc. IEEE*, 88(4):569–587, Apr. 2000.

[8] A. Kempf Y. Hao. On a non-fourier generalization of shannon sampling theory. *Proc. of Canadian Workshop on Information Theory*, pages 193–196, 2007.

[9] A. Kempf Y. Hao. On the stability of a generalized shannon sampling theorem. *Proc. of International Symposium on Information Theorem and its Applications*, Dec. 2008.

[10] A.I. Zayed. *Advances in Shannon's Sampling Theory*. CRC Press, Boca Baton, 1993.

Optimized Sampling Patterns for Practical Compressed MRI

Muhammad Usman(1) and Philip G. Batchelor(1)

(1) Division of Imaging Sciences, Kings College London, United Kingdom
muhammad.3.usman@kcl.ac.uk, philip.batchelor@kcl.ac.uk

Abstract:

The performance of compressed sensing (CS) algorithms is dependent on the sparsity level of the underlying signal, the type of sampling pattern used and the reconstruction method applied. The higher the incoherence of the sampling pattern used for under-sampling, less aliasing will be noticeable in the aliased signal space, resulting in better CS reconstruction. In this work, based on point spread function (PSF) properties, we compare random, Poisson disc and constrained random sampling patterns and show their usefulness in practical compressed sensing applied to dynamic cardiac magnetic resonance imaging (MRI).

Introduction

One of the main questions that arise in compressed sensing magnetic resonance imaging (CS-MRI) is: which type of sampling is optimal? The basic theory of compressed sensing as proposed by Donoho [1] and Candes [2] requires acquisition of randomized set of measurements. For MRI, this corresponds to the random sampling in Fourier domain (k-space) which results in incoherent aliasing artefacts in image space. However, random sampling requires bigger changes in amplitudes and polarity of MR system gradients, making it infeasible practically in an MR system.
Figure 1 shows one dimensional gradient variations for 2D random and uniform lattice sampling patterns. From the figure, it is evident that we have bigger changes in amplitude and polarity of gradients in case of random sampling pattern than uniform lattice. The solution to this problem is to use deterministic sampling patterns. The uniform lattice pattern is a deterministic pattern but yields coherent artefacts in its PSF and hence, it does not satisfy the basic requirements of compressed sensing theory. Our goal is to find deterministic sampling patterns that have incoherent artefacts in the PSF and can yield better CS reconstruction.

1. Candidate Sampling Patterns in CS

To have minimum aliasing due to sampling below the Nyquist rate, Nayak [3] defined the following properties of PSF of the ideal sampling pattern: The near zero region around the main lobe of the PSF should be as large as possible and outside that region, PSF should resemble white noise. The samples should be placed randomly but with a restricted maximum distance between samples. These two conditions are met by Poisson disc sampling [4]. Recently, Poisson disc sampling has been shown to give good results in parallel MRI due to better reconstruction conditioning [5]. However, it also has impractical gradient requirements. Gamper [6] defined constrained random pattern with incoherent artefacts in its PSF. Constrained random pattern is a normal lattice pattern with samples shifted along one dimension randomly by -1, 0 and +1. Hence, it is a normal lattice with constrained randomization added along one direction and has moderate gradient requirements. Figure 2 shows the three candidate sampling patterns (random, Poisson disc and constrained random) with the corresponding PSFs. Like Poisson disc sampling, the constrained random sampling has a near-zero region around the main lobe in its PSF and it also possesses the uniform density of sampling both locally and globally.

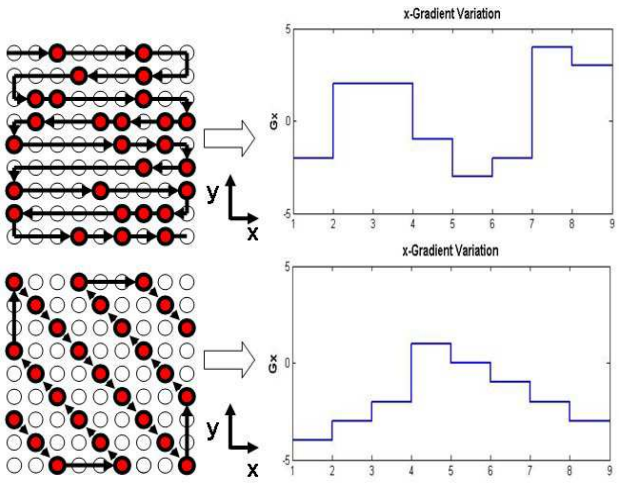


Figure 1: Gradient variation along one dimension in MR system for (a) random sampling (top) (b) uniform lattice sampling (bottom)

Additionally, due to added constrained randomization, the amplitudes of coherent side lobes in the PSF of constrained random pattern are also suppressed. In CS

recovery algorithms like OMP [6] which are based on picking the most significant component from the aliased space iteratively, the suppression of coherent artefacts ensures that only the right candidates are picked up in successive iterations.

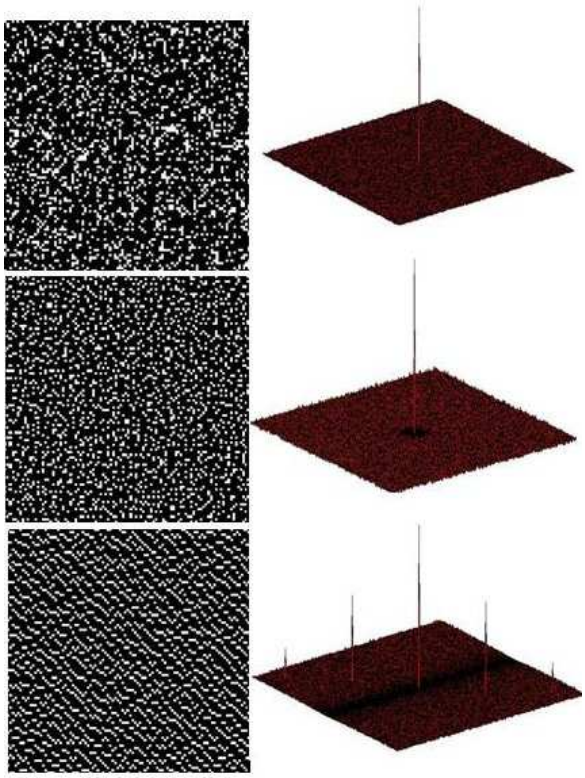


Figure 2: Three candidate sampling patterns and their corresponding PSFs: top to bottom: random, Poisson disc and constrained random

2. Experimental Setup

To test the performance of three sampling patterns in dynamic cardiac MRI, two sets of dynamic cardiac data of size (n_f x n_p x n_t, n_f: number of frequency encoding indices, n_p: number of phase encoding indices, n_t: number of time frames) (224x155x50) and (336x178x48) were acquired with a Philips MRI scanner 1.5 T, SSFP sequence, FOV 350x350 mm^2. We used the jittered grid approximation of the Poisson disc sampling as proposed by Cook [7]. For CS based reconstruction, the x-f space (x: spatial location, f: temporal frequency) is chosen to be the sparse representation [8]. The x-f space representation of the dynamic cardiac data is obtained by taking the Fourier transform of dynamic MR data along the temporal dimension. Figure 3 shows the dynamic cardiac MR data and its sparse representation.

For each frequency encoding index, the under-sampled data was simulated by applying the three sampling patterns to the fully sampled dynamic cardiac data in k_x-t space (k_x: phase encoding index, t: time) with varying acceleration factors/sampling factors (SF) from 3 to 7. The x-f space corresponding to each frequency encoding index was independently reconstructed by our modified OMP method with adaptive thresholding scheme [9]. The OMP algorithm stops when maximum residual aliasing intensity in x-f space reaches the intensity level of noise.

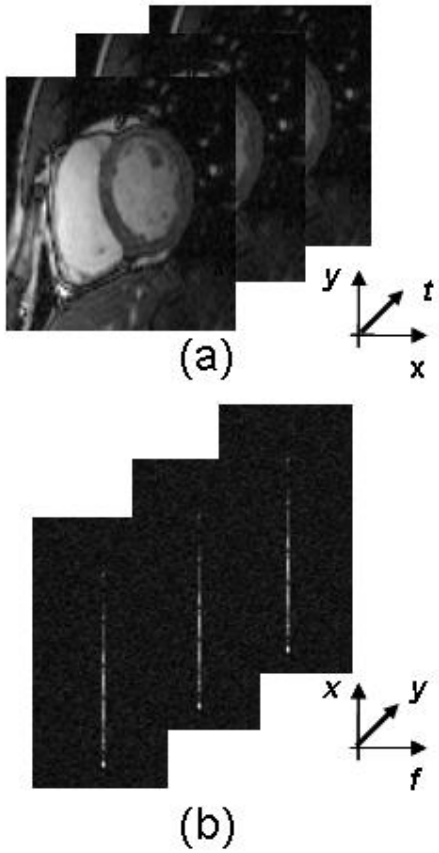


Figure 3: Dynamic cardiac MR data (a) and its x-f space representation (b), the frequency axis 'f' is centered around dc frequency (f=0)

3. Performance Results

The CS reconstruction results for candidate sampling patterns are shown in Figure 4 to Figure 9 with different acceleration factors. For the original cardiac frame shown in Figure 4 (a), the CS reconstruction results by OMP method with under-sampling factor (SF) of 3 are shown in Figure 4 (b), (c) and (d) for random, Poisson disc and constrained random sampling patterns. The corresponding temporal profiles are shown in Figure 5. The CS reconstruction results for acceleration factors of 5 and 7 are shown in Figure 6 to Figure 9. Up to the acceleration factor of 5, the CS reconstructed data has same spatial and temporal resolution for all three sampling patterns with nearly exact signal reconstruction achieved up to SF=3 (Figure 4 and Figure 5). Below SF=5, the temporal resolution of CS reconstructed data

with constrained random sampling gets worse than that for the other sampling patterns (Figure 9 d). This is due to the fact that with very high acceleration factors, many locations within the constrained random pattern will have zero probability of being picked up, as the sampling locations are constrained to only one sample shift from the uniform lattice grid.

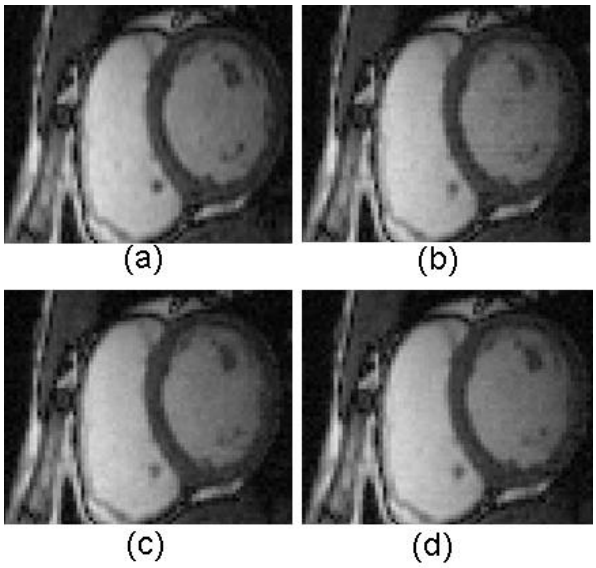


Figure 4: CS reconstruction results with SF=3:
(a) original cardiac frame, CS reconstructed data with (b) random sampling (c) Poisson disc sampling (d) constrained random sampling

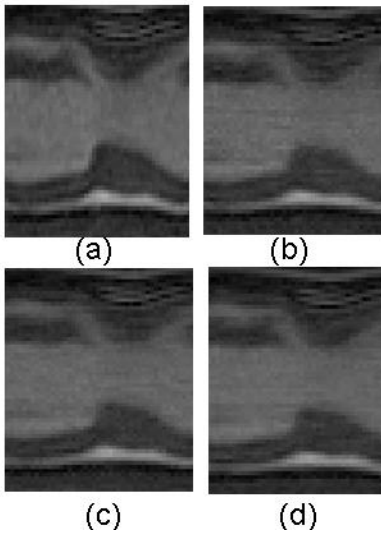


Figure 5:
CS reconstruction results with SF=3:
(a)original temporal profile, CS Reconstructed temporal profile with (b) random sampling
(c) Poisson disc sampling (d) constrained random sampling

4. Conclusion

We showed that the PSF properties of constrained random sampling are similar to the optimal sampling pattern (Poisson disc sampling). Up to the acceleration factor of 5, the quality of the reconstructed images and the temporal resolution of the CS reconstructed data are nearly the same for random, Poisson and constrained random sampling patterns. Since the constrained random sampling has moderate gradient requirements when compared with other optimal sampling schemes, it is an excellent choice to be used as an optimal sampling pattern in practical compressed MRI.

References:

[1] D.L.Donoho, “Compressed Sensing," IEEE transactions on information theory, vol.52, no. 4, pp. 1289-1306, 2006.
[2] E. Candes,”Compressive sampling," in Proceedings of the International Congress of Mathematicians, vol. 3, pp. 1433-1452, Madrid, Spain, 2006
[3] KS Nayak et al, “Randomized trajectories for reduced aliasing artifact”, in: Proceedings of the ISMRM, p 670., Sydney, 1998
[4] J. I. Yellot, “Spectral consequences of photoreceptor sampling in the rhesus retina.” Science 221, 382–385, 1985
[5] M. Lustig et al., “Autocalibrating Parallel Imaging Compressed Sensing using L1 SPIR-iT with Poisson-Disc Sampling and Joint Sparsity Constraints ISMRM Workshop on Data Sampling and Image Reconstruction, Sedona '09
[6] U. Gamper et al,”Compressed sensing in dynamic MRI,"MRM, vol. 59, no. 2, pp. 365-373, 2008.
[7] R. L. Cook, “ Stochastic sampling in computer graphics”, ACM Transactions on Graphics (TOG), vol.5, no. 1, p. 51-72, Jan 1986
[8] S. J. Malik et al, “x-f Choice: reconstruction of undersampled dynamic MRI by data-driven alias rejection applied to contrast-enhanced angiography. Stochastic sampling in computer graphics”, MRM, vol.56, p. 811-823, 2006
[9] M. Usman et al, “Adaptive thresholding scheme for OMP method applied to dynamic MRI”, Proc. ESMRMB, Valencia, vol. 25, no. 766, pp. 389, Oct 2008

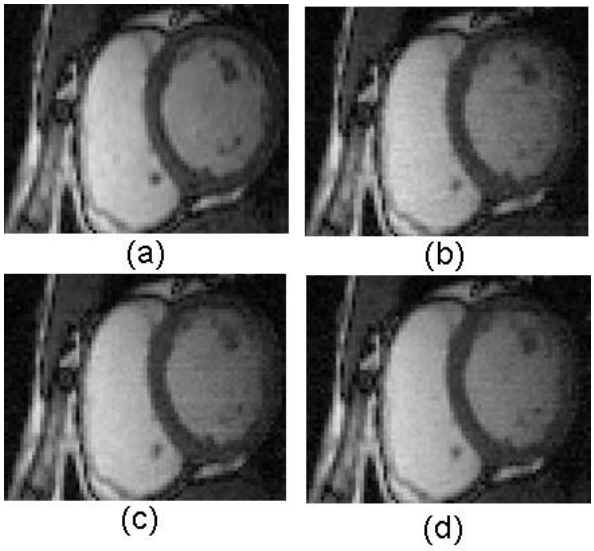


Figure 6: CS reconstruction results with SF=5:
(a) original cardiac frame, CS reconstructed data with (b) random sampling (c) Poisson disc sampling (d) constrained random sampling

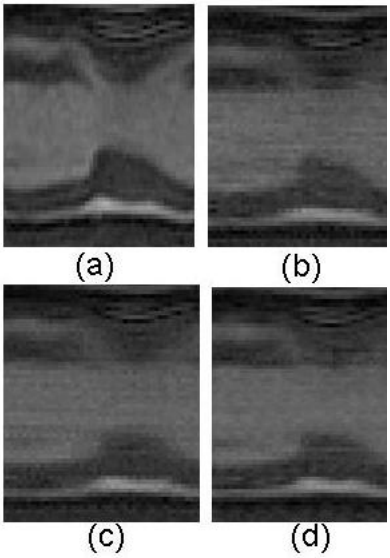


Figure 7:
CS reconstruction results with SF=5:
(a)original temporal profile, CS Reconstructed temporal profile with (b) random sampling
(c) Poisson disc sampling (d) constrained random sampling

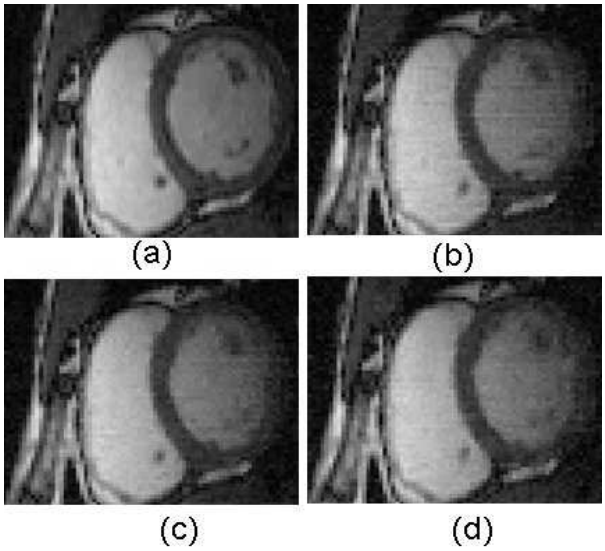


Figure 8: CS reconstruction results with SF=7:
(a) original cardiac frame, CS reconstructed data with (b) random sampling (c) Poisson disc sampling (d) constrained random sampling

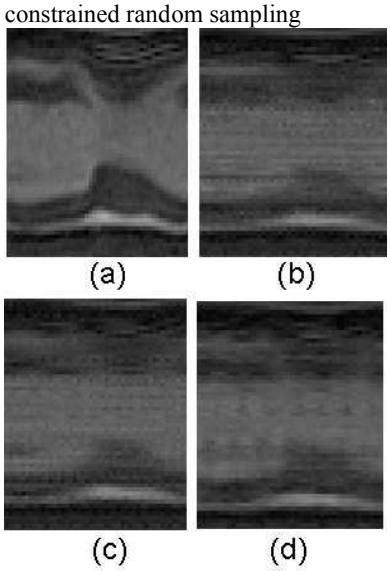


Figure 9:
CS reconstruction results with SF=7:
(a)original temporal profile, CS Reconstructed temporal profile with (b) random sampling
(c) Poisson disc sampling (d) constrained random sampling

A Study on Sparse Signal Reconstruction from Interlaced Samples by l1-Norm Minimization

Akira Hirabayashi (1)

(1) Yamaguchi University, 2-16-1, Tokiwadai, Ube City, Yamaguchi 755-8611, Japan.
a-hira@yamaguchi-u.ac.jp

Abstract:

We propose a sparse signal reconstruction algorithm from interlaced samples with unknown offset parameters based on the l_1-norm minimization principle. A typical application of the problem is superresolution from multiple low-resolution images. The algorithm first minimizes the l_1-norm of a vector that satisfies data constraint with the offset parameters fixed. Second, the minimum value is further minimized with respect to the parameters. Even though this is a heuristic approach, the computer simulations show that the proposed algorithm perfectly reconstructs sparse signals without failure when the reconstruction functions are polynomials and with more than 99% probability for large dimensional signals when the reconstruction functions are Fourier cosine basis functions.

1. Introduction

Sampling theory is at the interface of analog/digital conversion, and sampling theorems provide bridges between the continuous and the discrete-time worlds. A fundamental framework of the sampling theorems consists of data acquisition (sampling) process of a target signal and reconstruction process from the data. Classical studies assumed that both processes are fixed and known. Then, sampling theorems yield in linear formulations [9].

On the other hand, recent studies assume that sampling or reconstruction processes contain unknown factors. Then, sampling theorems become nonlinear. For example, Vetterli *et al.* discussed problems in which locations of reconstruction functions are unknown [11], [5]. They introduced the notion of rate of innovation, and provided perfect reconstruction procedures for signals with finite rate of innovation. The recent hot topic, compressive sampling, assumes that signals are sparse in the sense that signals are expressed by a small subset of reconstruction functions, but the subset is unknown [3], [1], [4]. It is interesting that the solution is obtained by the l_1-norm minimization.

In contrast to the above studies, problems with unknown factors in the sampling process have also been discussed. For example, sampling locations are assumed to be unknown and completely arbitrary in [8] and [2]. A more restricted sampling process is interlaced sampling [7], in which a signal is sampled by a sampling device several times with slightly shifted locations. If the offset parameters are unknown, the sampling theorem becomes nonlinear. A typical application is superresolution from a set of multiple low-resolution images. A replacement of a single high-rate A/D converter by multiple lower rate converters also yields within this formulation.

To this problem, Vandewalle *et al.* proposed perfect reconstruction algorithms under a condition that the total number of unknown parameters is less than or equal to the number of samples [10]. We can find, however, practical situations in which the condition is not true. The method proposed in [2] can be applied to such situations. However, it hardly provides a high quality stable result. In order to solve these difficulties, the present author proposed an algorithm that reconstructs the closest function to a mean signal under data constraint assuming that signals are generated from a probability distribution [6]. The mean signal is, however, not always available.

Hence, in this paper we propose a signal reconstruction algorithm from interlaced samples with unknown offsets using a relatively weak *a priori* knowledge, sparsity. The algorithm first minimizes the l_1-norm of a vector that satisfies data constraint with the offset parameters fixed. Then, the minimum value is further minimized with respect to the parameters. Even though this is a heuristic approach, the computer simulations show that the proposed algorithm perfectly reconstructs sparse signals without failure when the reconstruction functions are polynomials and with more than 99% probability for large dimensional signals when the reconstruction functions are Fourier cosine basis functions.

This paper is organized as follows. Section 2 formulates the fundamental framework and defines the notion of sparsity. Section 3 introduces interlaced sampling and summarizes the conventional studies. In Section 4, we propose the l_1-norm minimization algorithm. Section 5 evaluates the algorithm through simulations, and shows that the algorithm perfectly reconstruct sparse signals with high probability. Section 6 concludes the paper.

2. Sparse Signals

A signal f to be reconstructed is defined on a continuous domain $\mathcal{D}$. We assume that f belongs to a Hilbert space $H = H(\mathcal{D})$ of a finite dimension K. The inner product for f and g in H is denoted by $\langle f, g \rangle$, and the norm is induced as $\|f\| = \sqrt{\langle f, f \rangle}$. By using an arbitrarily fixed

basis $\{\varphi_k\}_{k=0}^{K-1}$, any f in H is expressed as

$$f = \sum_{k=0}^{K-1} a_k \varphi_k. \tag{1}$$

A K-dimensional vector with k-th element a_k is denoted by $\boldsymbol{a}$.

Definition 1 *A signal f is J-sparse if at most J coefficients of $\{a_k\}_{k=0}^{K-1}$ in Eq. (1) are non-zero and the rest are zero.*

It should be noted that unknown factors in J-sparse signals are not only values of non-zero coefficients but also their locations. Hence, there are $2J$ unknown factors in a J-sparse signal. If $2J \geq K$, then the number of unknown factors is more than K, which is the number of the original unknown coefficients $\{a_k\}_{k=0}^{K-1}$ without sparsity. Hence, in order for sparsity to be meaningful, we assume that

$$J < K/2.$$

In real applications, J is supposed to be much smaller than $K/2$.

3. Interlaced Sampling

Interlaced sampling means that a signal f is sampled M times by an identical observation device with offsets $\{\delta^{(m)}\}_{m=0}^{M-1}$, where $\delta^{(0)} = 0$. An M-dimensional vector with m-th element $\delta^{(m)}$ is denoted by $\boldsymbol{\delta}$. The observation device is characterized by sampling functions $\{\psi_n\}_{n=0}^{N-1}$, which are given *a priori*. Then, the sampling function for the n-th sample in the m-th sequence is given by

$$\psi_n^{(m)}(x) = \psi_n(x - \delta^{(m)}),$$

and the sample is expressed as

$$d_n^{(m)} = \langle f, \psi_n^{(m)} \rangle. \tag{2}$$

Let $\boldsymbol{d}$ be an MN-dimensional vector in which $d_n^{(m)}$ is the $n+mN$-th element. An $MN \times K$ matrix with the $n+mN$, k-th element $\langle \varphi_k, \psi_n^{(m)} \rangle$ is denoted by B_δ. Substituting Eq. (1) into Eq. (2) yields

$$B_\delta \boldsymbol{a} = \boldsymbol{d}. \tag{3}$$

For simplicity, we assume that the column vectors of B_δ are linearly independent. Figure 1 illustrates the formulation of interlaced sampling.

In order to reconstruct the signal f from interlaced samples with unknown offsets, we have to determine both $\{a_k\}_{k=0}^{K-1}$ and $\{\delta^{(m)}\}_{m=1}^{M-1}$. To this problem, Vandewalle *et al.* proposed perfect reconstruction algorithms under a condition that the number of unknown parameters is less than or equal to the number of samples $\{\{d_n^{(m)}\}_{n=0}^{N-1}\}_{m=0}^{M-1}$, or

$$K + M - 1 \leq MN. \tag{4}$$

We can find, however, practical situations in which the condition is not true. The method in [2] can be applied

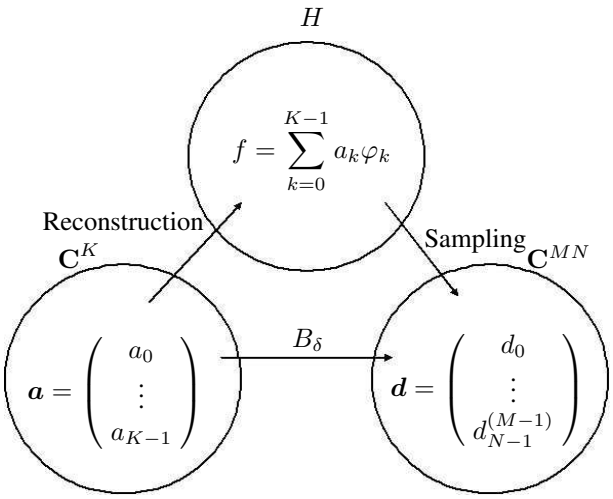


Figure 1: Formulation of sampling and reconstruction. The vector $\boldsymbol{a}$ is to be estimated from the vector $\boldsymbol{d}$. Note that there are unknown offset parameters $\boldsymbol{\delta}$ in B_δ.

to the situation without Eq. (4). However, the results obtained by the method tend to be unstable. The present author proposed an algorithm which uses a mean signal as a prior [6]. However, the mean signal is not always available. Hence, in this paper, we propose perfect reconstruction algorithms using a relatively weak prior, sparsity.

4. l_1-Norm Minimization Algorithm

The problem which we are going to solve in this paper is stated as follows.

Problem 1 *Determine J-sparse vector $\boldsymbol{a}$ and $\boldsymbol{\delta}$ which satisfy Eq. (3) under the condition that the column vectors of B_δ are linearly independent.*

Because of the linear independentness, a vector $\boldsymbol{a}$ that satisfies $B_\delta \boldsymbol{a} = \boldsymbol{d}$ is uniquely determined as

$$\boldsymbol{a} = B_\delta^\dagger \boldsymbol{d},$$

where $B_\delta^\dagger$ is the Moore-Penrose generalized inverse of B_δ. Let us define a matrix B_ε by setting an arbitrarily fixed parameter $\boldsymbol{\varepsilon}$ instead of $\boldsymbol{\delta}$. By using this matrix, a vector $\boldsymbol{c}_\varepsilon$ is defined as

$$\boldsymbol{c}_\varepsilon = B_\varepsilon^\dagger \boldsymbol{d}. \tag{5}$$

Then, our problem becomes a problem of finding a parameter $\boldsymbol{\varepsilon}$ such that the vector $\boldsymbol{c}_\varepsilon$ is J-sparse.

It is well-known that l_1-norm minimization is effective to promote sparsity as is used in the compressed sensing [3], [1], [4]. Hence, we also employ this principle to find J-sparse vector $\boldsymbol{c}_\varepsilon$. Now, our problem becomes the following problem.

Problem 2 *Determine $\boldsymbol{\varepsilon}$ that makes column vectors of the matrix B_ε linearly independent, and minimizes l_1-norm of $\boldsymbol{c}_\varepsilon$ in Eq. (5):*

$$\hat{\boldsymbol{\varepsilon}} = \mathrm{argmin}_{\boldsymbol{\varepsilon}} \, \|\boldsymbol{c}_\varepsilon\|_{l_1} = \mathrm{argmin}_{\boldsymbol{\varepsilon}} \, \|B_\varepsilon^\dagger \boldsymbol{d}\|_{l_1}. \tag{6}$$

Table 1: Parameters K, J, N and M used in simulations.

K	4	6	8	10	12
J	1	2	3	4	5
N	2	3	4	5	6
M	2	2	2	2	2

The solution to Problem 2 is different from that to Problem 1 in general. Similar to the compressed sensing, the former agrees with the latter in some cases. Theoretical analyses for the agreement are still under consideration. Instead, we show simulation results in this paper.

5. Simulations

We show computer simulations which demonstrate that the proposed algorithm perfectly reconstructs sparse signals under certain conditions. We consider two reconstruction functions, polynomial and Fourier cosine basis.

5.1 Polynomial reconstruction

Let H be a space spanned by functions

$$\varphi_k(x) = x^k \quad (0 \le k < K)$$

for $[0, l]$ where l is a positive real number. The inner product is defined by $\langle f, g \rangle = \frac{1}{l}\int_0^l f(x)\overline{g(x)}dx$. Sampling is assumed to be ideal, i.e., $d_n^{(m)} = f(x_n + \delta^{(m)})$. The sample point x_n is given by

$$x_n = \frac{(2n+1)l}{2N} \quad (n = 0, 1, \ldots, N-1),$$

which we call the base sequence. Let $l = N$ so that the sampling interval becomes one.

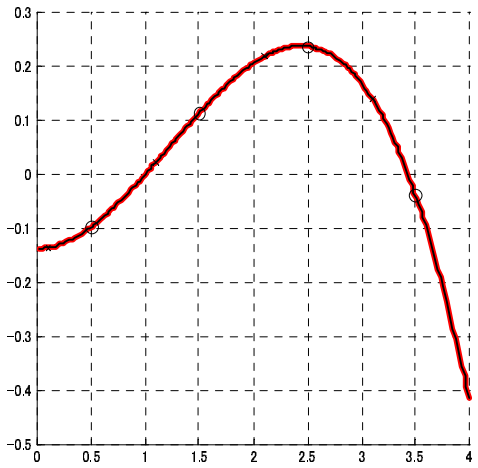
Figure 2 (a) shows a simulation result, in which the dimension of H is $K = 8$, sparsity parameter is $J = 3$, the number of samples in each sequence is $N = 4$, and the sequence was used $M = 2$ times. The offset parameter is $\delta^{(1)} = -0.4$. The black line shows the target signal f, and 'o' and 'x' respectively show the base and the first sequences. The red line shows the reconstructed signal, from which we can see the target signal is perfectly recovered. Figure 2 (b) shows that the l_1-norm of $\boldsymbol{c}_\varepsilon$ is indeed minimized at $\varepsilon = -0.4$. We repeated the simulation for one thousand target signals with the values shown in Table 1. Then, all of the signals are perfectly recovered as well as the offset parameters.

5.2 Fourier cosine basis reconstruction

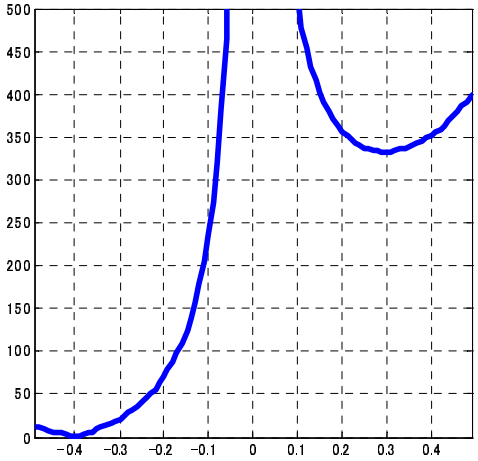
We used the same setup except that the reconstruction functions are

$$\varphi_k(x) = \begin{cases} 1 & (k = 0), \\ \sqrt{2}\cos\dfrac{k\pi x}{l} & (0 < k < K). \end{cases}$$

Under the above defined inner product, $\{\varphi_k\}_{k=0}^{K-1}$ is an orthonormal basis.



(a) Reconstruction result



(b) l_1-norm of $\boldsymbol{c}_\varepsilon$

Figure 2: Simulation result. The black line shows the target signal f, and 'o', 'x', and '+' respectively show the base, the first, and the second sequences. The red line shows the reconstructed signal which perfectly matches to the target signal.

Figure 3 (a) shows a simulation result, in which the dimension of H is $K = 60$, sparsity parameter is $J = 15$, the number of samples in each sequence is $N = 20$, and the sequence was used $M = 3$ times. The offset parameters are $\delta^{(1)} = -0.2$ and $\delta^{(2)} = 0.3$. The black line shows the target signal f, and 'o', 'x', and '+' respectively show the base, the first, and the second sequences. The red line shows the reconstructed signal, from which we can see the target signal is perfectly recovered.

Unfortunately, perfect reconstruction is not always achieved. Figure 4 shows failure rates [%] of perfect reconstruction with respect to K. The dotted red and the solid blue lines show the rates when $J = K/4$ and $J = K/6$, respectively. The failure rate for $J = K/4$ arrives at less than or equal to 1% when $K > 32$, while that for $J = K/6$ does so when $K > 30$.

Even though these results are only verified through simu-

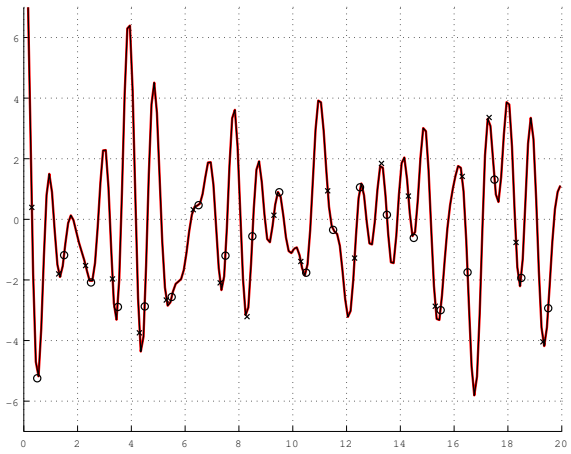


Figure 3: Simulation result for Fourier cosine basis functions. The black line shows the target signal f, and 'o', 'x', and '+' respectively show the base, the first, and the second sequences. The red line shows the reconstructed signal which perfectly matches to the target signal.

lations, the proposed approach is attractive because of its computational efficiency. It takes less than 0.4 second to find the solution for the case of $K = 60$, $N = 20$, and $M = 3$.

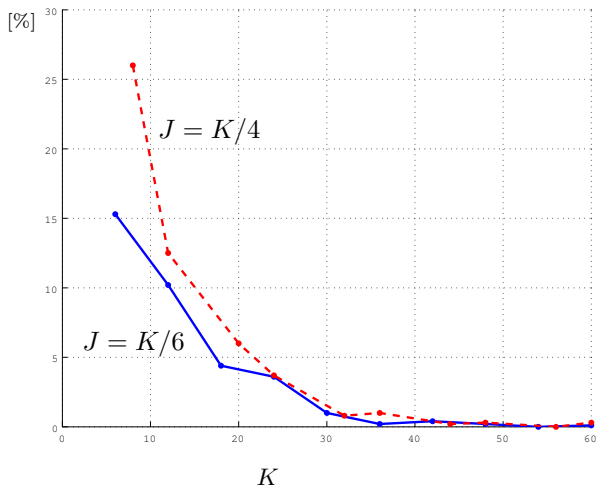


Figure 4: Failure rates of signal recovery when reconstruction functions are Fourier cosine basis functions.

6. Conclusion

We proposed a sparse signal reconstruction algorithm from interlaced samples with unknown offset parameters. The algorithm is based on the l_1-norm minimization principle: First, it minimizes the l_1-norm with the offset parameters fixed. Second, the minimum value is further minimized with respect to the parameters. Even though this is a heuristic approach, the computer simulations showed that the proposed algorithm perfectly reconstructs sparse signals without failure when the reconstruction functions are polynomials and with more than 99% probability for large dimensional signals when the reconstruction functions are Fourier cosine basis functions. Because of the computational efficiency, the proposed algorithm is very attractive. Theoretical analyses of these results are our most important future task.

Acknowledgment

This work was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), 20700164, 2008.

References:

[1] R.G. Baraniuk. Compressive sensing [lecture notes]. *IEEE Signal Processing Magazine*, 24(4):118–121, July 2007.

[2] J. Browning. Approximating signals from nonuniform continuous time samples at unknown locations. *IEEE Transactions on Signal Processing*, 55(4):1549–1554, April 2007.

[3] E.J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Thoery*, 52(2):489–509, February 2006.

[4] E.J. Candes and M.B. Wakin. An introduction to compressive sampling [a sensing/sampling paradigm that goes against the common knowledge in data acquisition]. *IEEE Signal Processing Magazine*, 25(2):21–30, March 2008.

[5] P. L. Dragotti, M. Vetterli, and T. Blu. Sampling moments and reconstructing signals of finite rate of innovation: Shannon meets Strang-Fix. *IEEE Transactions on Signal Processing*, 55(5):1741–1757, May 2007.

[6] Akira Hirabayashi and Laurent Condat. A study on interlaced sampling with unknown offsets. In *Proceedings of European Signal Processing Conference 2008 (EUSIPCO2008)*, volume CD-ROM, 2008.

[7] R.J. Marks II. *Introduction to Shannon Sampling and Interpolation Theory*. Springer-Verlag, New York, 1991.

[8] P. Marziliano and M. Vetterli. Reconstruction of irregularly sampled discrete-time bandlimited signals with unknown sampling locations. *IEEE Transactions on Signal Processing*, 48(12):3462–3471, December 2000.

[9] M. Unser. Sampling—50 Years after Shannon. *Proceedings of the IEEE*, 88(4):569–587, April 2000.

[10] P. Vandewalle, L. Sbaiz, J. Vandewalle, and M. Vetterli. Super-resolution from unregistered and totally aliased signals using subspace methods. *IEEE Transactions on Signal Processing*, 55(7):3687–3703, July 2007.

[11] M. Vetterli, P. Marziliano, and T. Blu. Sampling signals with finite rate of innovation. *IEEE Transactions on Signal Processing*, 50(6):1417–1428, June 2002.

Multiresolution analysis on multidimensional dyadic grids

Douglas A. Castro[(1)], Sônia M. Gomes[(1)], Anamaria Gomide[(2)], Andrielber S. Oliveira[(1)], Jorge Stolfi[(2)]

(1) IMECC-Unicamp, Caixa Postal 6065, CEP 13083-859 Campinas-SP, Brazil.
(2) IC-Unicamp, Caixa Postal 6176, CEP 13081-970 Campinas-SP, Brazil.
{douglas,andriel,soniag}@ime.unicamp.br {anamaria,stolfi}@ic.unicamp.br

Abstract:

We propose a modified adaptive multiresolution scheme for representing d-dimensional signals which is based on cell-average discretization in dyadic grids. A dyadic grid is an hierarchy of meshes where a cell at a certain level is partitioned into two equal children at the next refined level by hyperplanes perpendicular to one of the coordinate axes which varies cyclically from level to level. Adaptivity is obtained by interrupting the refinement at the locations where appropriate scale (wavelet) coefficients are sufficiently small. One important aspect of such multiresolution representation is that we can use a binary tree data structure in all dimensions, that helps to compress data while still being able to navigate through it. Dyadic grids provide a more gradual refinement as compared with traditional multiresolution analyses that use, for instance, different quad-trees or oct-trees in 2D or 3D multiresolution applications. The cells may have different scales in different directions, this property can be explored to improve data compression of signals having anisotropic aspects.

1. Introduction

In recent years, many multiscale techniques have been used to provide more efficient algorithms than those that use just one level of resolution. In such frameworks, the differences between the information at consecutive levels of refinement are computed, and only the significant coefficients are stored. These are the principles of wavelet compression which have being successfully applied in many different contexts [3]. For example, multiresolution finite volume schemes of Müller [6] and Domingues et al. [4] use adaptive grids that are dynamically obtained by taking local regularity information indicated by wavelet coefficients in the context of multiresolution analysis for cell averages of signals. Such adaptive discretizations allow the efficient solution of problems with vastly different scales of detail in different parts of the domain.

For computational efficiency, one important aspect of such multiresolution methods is the topology of the mesh and data structure used to represent it. Often quad-grids and oct-grids are used for 2D and 3D domains, respectively, represented by quad-tree and oct-tree data structures [1]. We describe here a type of mesh, the *dyadic grid*, that can be efficiently represented by a binary tree, in domains of arbitrary dimension. For illustration, we apply adaptive dyadic grids to multiresolution analysis, using cell averaging as the discretization method.

The paper is organized as follows. In Section 2 we define dyadic grids and related concepts. In Section 3 we present a general overview of multiresolution analysis. Section 4 contains numerical results on sample problems to show the efficiency of the proposed scheme.

2. Dyadic grids

Let the coordinates of $\mathbb{R}^d$ be indexed from 0 to $d-1$. An infinite *dyadic grid* is a hierarchy of meshes that begins with a d-cube at level $k = 0$, and, for each higher level $k > 0$, is the result of dividing each cell of level k into two equal children by a hyperplane perpendicular to the coordinate axis $(k \bmod d)$ [2]. Figure 1 illustrates five steps of the refinement process for $d = 3$.

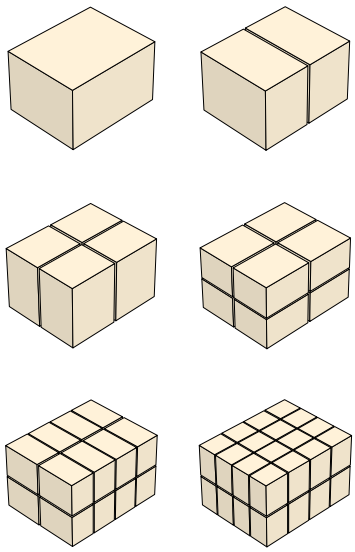


Figure 1: 3D dyadic grids.

In practice, one uses only finite segments of this grid, where the subdivision stops at a maximum level. In a *regular* dyadic grid, the refinement stops at the same level everywhere. In an *irregular* grid, the maximum level varies from place to place.

The topology of a dyadic grid can be represented by a *0-2 binary tree*. This is a data structure consisting of a set

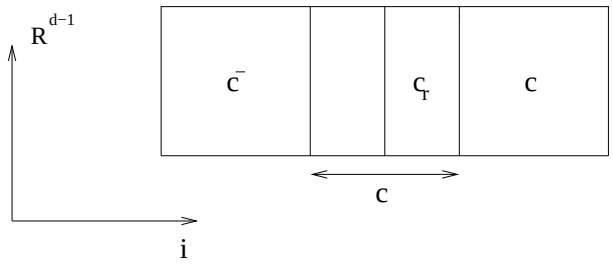


Figure 2: Definition of c^-, c^+, c, c_r.

of elements named *nodes*, among which there is a special node r, the *root*; every node has either zero or two children nodes; and every node, except the root, has exactly one parent node. A node that has no children is called a *leaf node*. The children of a non-leaf node t are called the *left child* t_ℓ and the *right child* t_r.
Each node of this tree represents a cell that appeared at some level of the subdivision; the leaf nodes represent the cells that weren't divided. By convention, the left child t_ℓ of a non-leaf node t in level k represents the "lower" half c_ℓ of the cell c represented by t; that is, the half whose projection on the axis $i = k \bmod d$ has smallest i-coordinates.

3. Multiresolution analysis

In mutiresolution analysis, signals can be represented in two ways, as ordinary samples at each scale, or as differences between two consecutive scales. Connecting these two views are the *prediction* and the *restriction* operators. The prediction operator P_k^{k+1} takes information from a coarse level k and gives an estimate for the information at the next finer level $k+1$. Conversely, the restriction operator P_{k+1}^k takes information from a fine level $k+1$ and gives an estimate of the information at a coarser level k.
In this paper, the samples of a $d-$dimensional signal f are averages computed over the cells of a d-dimensional dyadic grid. That is, the sample associated with a cell c in level k of the grid is

$$\bar{f}_c^k = \frac{1}{|c|}\int_c f(x)\mathrm{dx}, \tag{1}$$

where $|c|$ is the volume of c. The restriction operation is therefore (trivially and exactly) the sum of the averages in the children cells,

$$\bar{f}_c^k = \frac{1}{2}\Big[\bar{f}_{c_\ell}^{k+1} + \bar{f}_{c_r}^{k+1}\Big]. \tag{2}$$

In the other direction, we predict the cell average of a child cell c_r or c_ℓ by the formulas

$$\bar{f}_{c_r}^{k+1} \approx \hat{f}_{c_r}^{k+1} \quad = \quad \bar{f}_c^k + \frac{1}{8}\Big[\bar{f}_{c^+}^k - \bar{f}_{c^-}^k\Big] \tag{3}$$

$$\bar{f}_{c_\ell}^{k+1} \approx \hat{f}_{c_\ell}^{k+1} \quad = \quad \bar{f}_c^k - \frac{1}{8}\Big[\bar{f}_{c^+}^k - \bar{f}_{c^-}^k\Big] \tag{4}$$

where c^- and c^+ are the two closest neighbor cells of c at level k in the direction of refinement. See Figure 2. These estimators are exact for quadratic polynomials.

Detail coefficients. In the structure we do not store the averages ($\bar{f}_c^k$ or $\hat{f}_c^k$), but only the *details* or *wavelet coeficients*. Each detail d_c^k is the difference between the exact average in the cell c and the value predicted for it by formulas (3) and (4) from the cell's parent and its neighbors:

$$d_c^{k+1} = \bar{f}_c^{k+1} - \hat{f}_c^{k+1}. \tag{5}$$

Note that the detail of the root cell is not defined.

Analysis and synthesis. The *analysis algorithm* computes the details of every cell, given the average values $\bar{f}_c^k$ for every cell c. It scans the tree bottom-up, level by level. For each non-leaf cell c in level k, it executes

$$\begin{array}{l} \bar{\delta} \leftarrow \frac{1}{2}\Big[\bar{f}_{c_r}^{k+1} - \bar{f}_{c_\ell}^{k+1}\Big]; \\ \hat{\delta} \leftarrow \frac{1}{8}\Big[\bar{f}_{c^+}^{k} - \bar{f}_{c^-}^{k}\Big]; \\ \delta \leftarrow \bar{\delta} - \hat{\delta} \\ d_{c_r}^k \leftarrow +\delta; \\ d_{c_\ell}^k \leftarrow -\delta. \end{array} \tag{6}$$

Once the detail d_c^k of a cell has been computed, its average $\bar{f}_c^k$ is no longer needed, so we may store the detail in its place. In the root node r, however, we must still keep the average $\bar{f}_r^0$ of the function over the whole domain.
The inverse of the analysis algorithm is the *syntesis algorithm*, which recomputes the averages $\bar{f}_c^k$ from the details. It scans the tree top down, level by level. At each cell c in level k, it executes

$$\begin{array}{l} \hat{\delta} \leftarrow \frac{1}{8}\Big[\bar{f}_{c^+}^{k} - \bar{f}_{c^-}^{k}\Big]; \\ \bar{\delta} \leftarrow \hat{\delta} - d_{c_r}^{k+1}; \\ \bar{f}_{c_r}^{k+1} = \bar{f}_c^k + \bar{\delta} \\ \bar{f}_{c_\ell}^{k+1} = \bar{f}_c^k - \bar{\delta}. \end{array} \tag{7}$$

After this step, the details $d_{c_r}^{k+1}$ and $d_{c_\ell}^{k+1}$ of the children are no longer needed, and can be overwritten with the reconstructed averages $\bar{f}_{c_r}^{k+1}$ and $\bar{f}_{c_\ell}^{k+1}$.

Compact representation. These algorithms show that knowledge of the cell averages for all leaves is equivalent to knowledge of the average value $\bar{f}_r^0$ for the root cell together with the detail of every right child cell. To save space, we could store the detail of the right child in its parent's node (and keep the domain average $\bar{f}_r^0$ in variable external to the tree). Then the leaf nodes would carry no information, and could be omitted from the structure. We will refer to this variant (which is an ordinary binary tree) as the *compact* tree representation.

Adaptive resolution grid. As in any wavelet representation, we can save space and processing time by pruning all sub-trees which do not contribute significantly to the reconstructed signal. If we start with a tree of sufficient depth, we can eliminate all sibling leaf nodes c_ℓ and c_r such that $\left|d_c^k\right|$ falls below a prescribed tolerance ϵ_k. This condition implies that the predictions $\hat{f}_{c_\ell}^{k+1}$ and $\hat{f}_{c_r}^{k+1}$ will be very close to the actual averages $\bar{f}_{c_\ell}^{k+1}$ and $\bar{f}_{c_r}^{k+1}$. Here we use Harten's thresholds [5],

$$\epsilon_k = (1-q)q^{(L-k)}\epsilon, \tag{8}$$

where q and ϵ are specified by the user, with $\epsilon > 0$ and $0 < q < 1$, and L is the maximum level of the initial tree.

4. Numerical results

In order to compare the efficiencies of dyadic grids and quad-grids, we performed the multiresolution analyses of two different examples in 2D, using cell-average discretization. In all tests, the root cell was the rectangle $[0,1] \times [0, \frac{\sqrt{2}}{2}]$, and the starting tree was a uniform grid with $2^{10} \times 2^{10} = 2^{20} = 1,048,576$ leaf cells. This corresponds to tree structures with $L = 20$ and $L = 10$ levels for dyadic grid and quad-grid frameworks, respectively. The cell averages $\bar{f}_c^L$ were computed for every leaf cell c by Gaussian quadrature with 5×5 sampling points. The trees were pruned as described in the previous section, with threshold parameters $\epsilon = 0.1$ and the $q = 0.5$. The number of non-leaf nodes in the initial tree was $\sum_{i=0}^{19} 2^i = 1,048,575$ for the dyadic grid, and $\sum_{i=0}^{9} 4^i = 349,525$ for the quad-grid.
In the first test, we used the signals

$$f(x,y) = 1 - \tanh(100(x - 0.2 - t) + 0.001(y-1)), \quad (9)$$

for t varying from 0 to 0.6 in steps of 0.1. Equation (9) describes a 2D smooth step function with an almost vertical straight front, moving form left to right. Figure 3 shows the dyadic grid and the corresponding tree at $t = 0.1$, after pruning cells with small details. Figure 4 shows the corresponding quad-grid and quad-tree. Figure 5 shows the number of leaf cells in both grids for each time step, as a percentage of the number of leaves in the uniform grid.
For the second test, we used the signals

$$f(x,y) = \begin{cases} 1 \text{ if } \sqrt{(x-0.5)^2 + (y-0.35)^2} < t, \\ 0 \text{ otherwise.} \end{cases} \quad (10)$$

for t varying between 0.05 and 0.35 in steps of 0.05. Equation (10) describes a step function with a sharp circular front, expanding from the center of the domain. Figure 6 shows the dyadic grid and its tree at $t = 0.2$, and Figure 7 shows the corresponding quad-grid and its quad-tree. The number of leaves is plotted in Figure 8.

Space efficiency. If leaves are explicitly represented in the tree structure, and all nodes have the same fields, then the space E used by the structure is $E = (pA + B)n$, where n is the number of nodes, p is the number of pointers in each node, A is the size of a pointer in bytes, and B is the size of any additional information stored in each node (such as the detail coefficients d_c^k). In all these trees we have $n = (pm - 1)/(p - 1) \approx mp/(p - 1)$, where m is the number of leaf nodes.
From the plots in Figure 5, we see that, in the first test, the quad-grid ($p = 4$) had about 8 times as many leaf cells as the dyadic grid ($p = 2$), and therefore about 5 times as many tree nodes, for the same accuracy. Assuming $A = 4$ and $B = 8$ bytes, we conclude that the quad-tree used $5(24/16) \approx 7.5$ times as much storage as the dyadic grid.

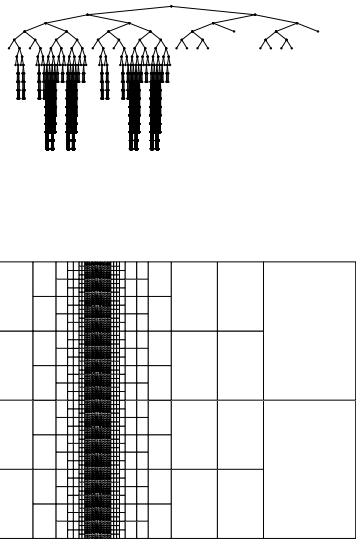


Figure 3: Pruned dyadic tree (top) and dyadic grid (bottom) for the first signal at $t = 0.1$.

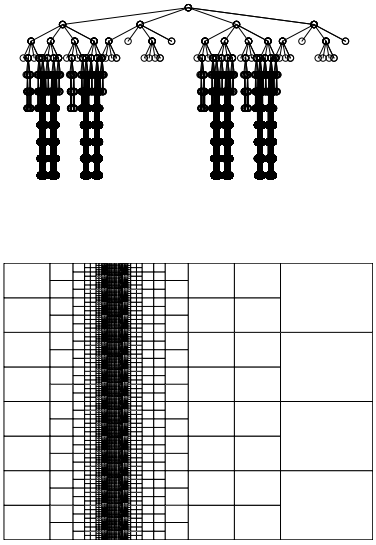


Figure 4: Pruned quad-tree (top) and quad-grid (bottom) for the first signal at $t = 0.1$..

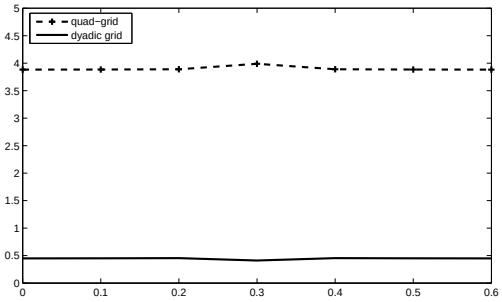


Figure 5: Leaf count in the pruned trees for the first test..

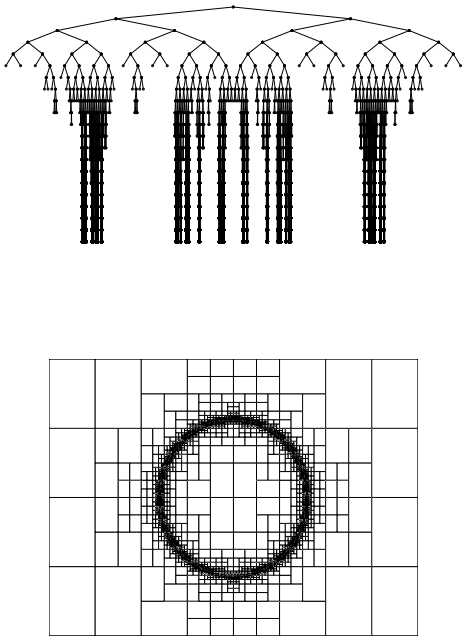


Figure 6: Pruned dyadic tree (top) and dyadic grid (bottom) for the second signal at $t = 0.2$..

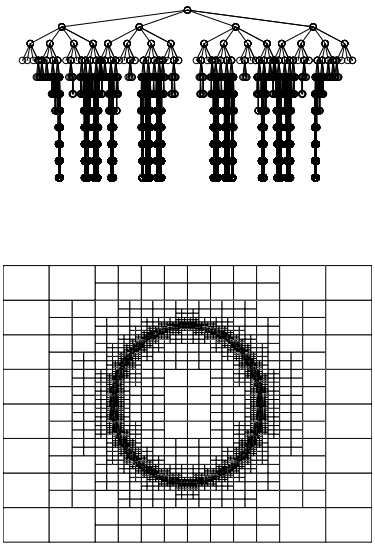


Figure 7: Pruned quad-tree (top) and quad-grid (bottom) for the second signal at $t = 0.2$..

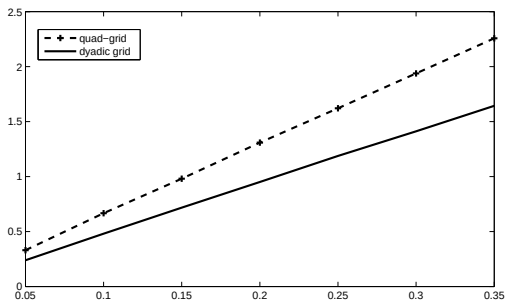


Figure 8: Leaf count in the pruned trees for the second test..

In the second test, the quad-grid had about 1.32 times as many leaf cells as the dyadic grid, and therefore about 0.88 as many tree nodes. With the same A and B, the quad grid still used $0.88(24/16) \approx 1.32$ times as much space as the dyadic grid.
Had we used the compact representation of the tree, with omitted leaves, the storage cost would be $E = (pA + (p-1)B)(n-m)$. The quad-tree would use 7.5 times as much storage as the dyadic tree in the first example, and 1.1 times as much in the second example.

5. Conclusions

Our tests show that adaptive dyadic grids are substantially more efficient than quad-grids for the same level of accuracy, both in terms of space needed to store the topology (tree structure) of the grid, and in the number of leaf cells retained — which determines the time cost of most adaptive numeric algorithms.

6. Acknowledgments

The authors thank CNPq (grants 06631/07-5, 472402/07-2, and 142191/06-0) and FAPESP (07/52015-0) for financial support.

References:

[1] B. L. Bihari and A. Harten. Multiresolution schemes for the numerical solution of 2-d conservation laws. *SIAM J. Sci. Comput.*, 18(2):315–354, 1997.

[2] C. G. S. Cardoso, M. C. Cunha, A. Gomide, D. J. Schiozer, and J. Stolfi. Finite elements on dyadic grids with applications. *Mathematics and Computers in Simulation*, 73:87–104, 2006.

[3] A. Cohen. *Wavelet Methods in Numerical Analysis. Handbook of Numerical Analysis.* in: Ph. Ciarlet and J. L. Lions (Eds.), Handbook of Numerical Analysis, Vol VII, Elsevier, Amsterdam, 2000.

[4] M. O. Domingues, S. M. Gomes, O. Roussel, and K. Schneider. An adaptive multiresolution scheme with local time-stepping for evolutionary pdes. *Journal of Computational Physics*, 227:3758–3780, 2008.

[5] A. Harten. Multiresolution representation of cell-averaged data. Technical Report CAM/Report/94-21, UCLA, Los Angeles, US, July 1994.

[6] S. Muller. *Adaptive Multiscale Schemes for Conservation Laws*. Vol. 27 of Lecture Notes in Computational Science and Engineering, Springer, Heidelberg, 2003.

Adaptive and Ultra-Wideband Sampling via Signal Segmentation and Projection

Stephen D. Casey [(1)], Brian M. Sadler[(2)]

(1) Department of Mathematics and Statistics, American University, Washington, DC, USA .
(2) Army Research Laboratory, Adelphi, MD, USA.
scasey@american.edu, bsadler@arl.army.mil

Abstract:

Adaptive frequency band (AFB) and ultra-wide-band (UWB) systems require either rapidly changing or very high sampling rates. Conventional analog-to-digital devices have non-adaptive and limited dynamic range. We investigate AFB and UWB sampling via a basis projection method. The method decomposes the signal into a basis over time segments via a continuous-time inner product operation and then samples the basis coefficients in parallel. The signal may then be reconstructed from the basis coefficients to recover the signal in the time domain. We develop the procedure of this method, analyze various methods for signal segmentation and close by creating systems designed for binary signals.

1. Introduction

Adaptive frequency band (AFB) and ultra-wide-band (UWB) systems, requiring either rapidly changing or very high sampling rates, stress classical sampling approaches. At UWB rates, conventional analog-to-digital devices have limited dynamic range and exhibit undesired nonlinear effects such as timing jitter. Increased sampling speed leads to less accurate devices that have lower precision in numerical representation. This motivates alternative sampling schemes that use mixed-signal approaches, coupling analog processing with parallel sampling, to provide improved sampling accuracy and parallel data streams amenable to lower speed (parallel) digital computation.

We investigate AFB and UWB sampling via a basis projection method. The method was introduced as a means of UWB parallel sampling by Hoyos *et. al.* [7] and applied to UWB communications systems [8, 9, 10]. The method first decomposes the signal into a basis over time segments via a continuous-time inner product operation and then samples the basis coefficients in parallel. The signal may then be reconstructed from the basis coefficients to recover time domain samples, or further processing may be carried out in the new domain [7].

We address several issues associated with the basis-expansion and sampling procedure, including choice of basis, truncation error, rate of convergence and segmentation of the signal. We develop a mathematical model of the procedure, using both standard (sine, cosine) basis elements and general basis elements, and give this representation in both the time and frequency domains. We compute exact truncation error bounds, and compare the method with traditional sampling. We close by developing the method for binary signals.

2. Sampling via Projection

Let f be a signal of finite energy whose Fourier transform $\widehat{f}$ has compact support, i.e., $f, \widehat{f} \in L^2$, with $\mathrm{supp}(\widehat{f}) \subset [-\Omega, \Omega]$. The signal is in the Paley-Wiener class $PW(\Omega)$. For a block of time T_c, let

$$f(t) = \sum_{k\in\mathbb{Z}} f(t)\chi_{[(k)T_c,(k+1)T_c]}(t)\,.$$

For this original development, we keep T_c fixed. We later let T_c be adaptive and will denote the adaptive time segmentation as $\tau_c(t)$. A given block $f_k(t) = f(t)\chi_{[(k)T_c,(k+1)T_c]}(t)$ can be T_c- periodically continued, getting

$$(f_k)^\circ(t) = (f(t)\chi_{[(k)T_c,(k+1)T_c]}(t))^\circ\,.$$

Expanding $(f_k)^\circ(t)$ in a Fourier series, we get

$$(f_k)^\circ(t) = \sum_{n\in\mathbb{Z}} \widehat{(f_k)^\circ}[n]e^{(2\pi int/T_c)}\,,\ \text{where}$$

$$\widehat{(f_k)^\circ}[n] = \frac{1}{T_c}\int_{(k)_{T_c}}^{(k+1)T_c} f(t)e^{(-2\pi int/T_c)}\,dt\,.$$

Given that the original function f is Ω band-limited, we can estimate the value of n for which $f_k[n]$ is non-zero. At minimum, $f_k[n]$ is non-zero if

$$\frac{n}{T_c} \le \Omega\,,\text{or equivalently, } n \le T_c\cdot\Omega\,.$$

Let

$$N = \lceil T_c\cdot\Omega\rceil\,.$$

(Note that the truncated block functions f_k are not band-limited. We discuss this in section 3.) For this choice of N, we compute

$$\begin{aligned} f(t) &= \sum_{k\in\mathbb{Z}} f(t)\chi_{[(k)T_c,(k+1)T_c]}(t) \\ &= \sum_{k\in\mathbb{Z}}\Big[(f_k)^\circ(t)\Big]\chi_{[(k)T_c,(k+1)T_c]}(t) \\ &\approx \sum_{k\in\mathbb{Z}}\Big[\sum_{n=-N}^{n=N}\widehat{(f_k)^\circ}[n]e^{(2\pi int/T_c)}\Big]\chi_{[(k)T_c,(k+1)T_c]}(t)\,. \end{aligned}$$

Given this choice of the standard (sines, cosines) basis, we can, for a fixed value of N, adjust to a large bandwidth Ω by choosing small time blocks T_c. Also, after a given set of time blocks, we can deal with a increase or decrease in bandwidth Ω by again adjusting the time blocks, e.g., given an increase in Ω, decrease the time blocks adaptively to $\tau_c(t)$, and vice versa. There is, of course, a price to be paid. The quality of the signal, as expressed in the accuracy the representation of f, depends on N, Ω and T_c.

Theorem : [The Projection Formula] Let f, $\widehat{f} \in L^2(\mathbb{R})$ and $f \in PW_\Omega$, i.e. $\mathrm{supp}(\widehat{f}) \subset [-\Omega, \Omega]$. Let T_c be a fixed block of time. Then, for $N = \lceil T_c \cdot \Omega \rceil$, $f(t) \approx f_{\mathcal{P}}(t)$, where

$$f_{\mathcal{P}}(t) = \sum_{k\in\mathbb{Z}} \left[\sum_{n=-N}^{N} f_k[n] e^{(2\pi i n t/T_c)} \right] \chi_{[kT_c,(k+1)T_c]}(t). \tag{1}$$

The Projection Method can adapt to changes in the signal. Suppose that the signal $f(t)$ has a band-limit $\Omega(t)$ which changes with time. For example, let f be a signal from a cell phone which changes from voice to a highly detailed musical piece. This change effects the time blocking $\tau_c(t)$ and the number of basis elements $N(t)$. This, of course, makes the analysis more complicated, but is at the heart of the advantage the Projection Method has over conventional methods.

During a given $\tau_c(t)$, let $\overline{\Omega}(t) = \sup\{\Omega(t) : t \in \tau_c(t)\}$. For a signal f that is $\Omega(t)$ band-limited, we can estimate the value of n for which $f_k[n]$ is non-zero. At minimum, $f_k[n]$ is non-zero if

$$\frac{n}{\tau_c(t)} \leq \overline{\Omega}(t)\,, \text{or equivalently, } n \leq \tau_c(t) \cdot \overline{\Omega}(t)\,.$$

Let

$$N(t) = \lceil \tau_c(t) \cdot \overline{\Omega}(t) \rceil\,.$$

For this choice of $N(t)$, we have the following.

Theorem : [The Adaptive Projection Formula] Let f, $\widehat{f} \in L^2(\mathbb{R})$ and f have a variable but bounded band-limit $\Omega(t)$. Let $\tau_c(t)$ be an adaptive block of time and given $\tau_c(t)$, let $\overline{\Omega}(t) = \sup\{\Omega(t) : t \in \tau_c(t)\}$. Then, for $N(t) = \lceil \tau_c(t) \cdot \Omega(t) \rceil$, $f(t) \approx f_{\mathcal{P}}(t)$, where

$$f_{\mathcal{P}}(t) = \sum_{k\in\mathbb{Z}} \left[\sum_{n=-N(t)}^{N(t)} f_k[n] e^{(2\pi i n t/\tau_c)} \right] \chi_{[k\tau_c,(k+1)\tau_c]}(t). \tag{2}$$

In comparison, Shannon Sampling examines the function at specific points, then uses those individual points to recreate the signal. The Projection Method breaks the signal into segments in the time domain and then approximates their respective periodic expansions with a Fourier series. This process allows the system to individually evaluate each piece and base its calculation on the needed bandwidth. The individual Fourier series are then summed, recreating a close approximation of the original signal. It is important to note that instead of fixing T_c, the method allows us to fix any of the three while allowing the other two to fluctuate. The easiest and most practical parameter from the design factor to fix is N. For situations in which the bandwidth does not need flexibility, it is possible to fix Ω and T_c by the equation $N = \lceil T_c \cdot \Omega \rceil$. However, if greater bandwidth Ω is need, choose shorter time blocks T_c.

The Projection Method adapts to general orthonormal systems, much as Kramer-Weiss extends sampling to general orthonormal bases. Given a function f such that $f \in PW_\Omega$, let T_c be a fixed time block. Define $f(t)$, $f_k(t)$ and ${f_k}^\circ(t)$ as in the beginning of the computation above. Now, let $\{\varphi_n\}$ be a general orthonormal system for $L^2[0, T_c]$. Then,

$${f_k}^\circ(t) = \sum_{n=-\infty}^{\infty} f_k[n]\varphi_n(t), \text{ where } f_k[n] = \langle {f_k}^\circ, \varphi_n \rangle.$$

Since $f \in PW_\Omega$, there exists $N = N(T_c, \Omega)$ such that $f_k[n] = 0$ for all $n > N$. Therefore, $f(t) \approx f_{\mathcal{P}}(t)$, where

$$f_{\mathcal{P}}(t) = \sum_{k=-\infty}^{\infty} \left[\sum_{n=-N}^{N} f_k[n]\varphi_n(t) \right] \chi_{[kT_c,(k+1)T_c]}(t). \tag{3}$$

Given characteristics of the input class signals, the choice of basis functions used in the the Projection Method can be tailored to optimal representation of the signal or a desired characteristic in the signal. We develop a Walsh system for binary signals in section 4.

We close this section with a different system of segmentation for the time domain. This was created because it is relatively easy to implement, cuts down on frequency error and has no loss of data in time. It was developed by studying the de la Vallée-Poussin kernel used in Fourier series. Let $0 < r < T_c/2$ and let

$$\mathrm{Tri}_L(t) = \max\{[((T_c/(4r)) + r) - |t|/(2r)], 0\}\,,$$

$$\mathrm{Tri}_S(t) = \max\{[((T_c/(4r)) + r - 1) - |t|/(2r)], 0\}$$

and

$$\mathrm{Trap}(t) = \mathrm{Tri}_L(t) - \mathrm{Tri}_S(t)\,.$$

The Trap function has perfect overlay in the time domain and $1/\omega^2$ decay in frequency space. When one time block is ramping down, the adjacent block is ramping up at exactly the same rate. This leads to the Projection formula

$$\sum_{k\in\mathbb{Z}} \left[\sum_{n=-N}^{N} ((f \cdot \mathrm{Trap})_k[n] e^{(2\pi i n t/(T_c+r))} \right] \mathrm{Trap}(t - k(T_c/2))\,.$$

3. Error Analysis

To compute truncation error, we first calculate the Fourier transform of both sides of the equation. Let $f \in PW(\Omega)$, so $f \in L^2$ and Ω band-limited. For $N = \lceil T_c \cdot \Omega \rceil$,

$$f_{\mathcal{P}}(t) = \sum_{k\in\mathbb{Z}} \left[\sum_{n=-N}^{N} f_k[n] e^{(2\pi i n t/T_c)} \right] \chi_{[kT_c,(k+1)T_c]}(t)$$

Taking the transform of both sides and evoking the relationship between the transform and convolution gives

$$\widehat{f_{\mathcal{P}}}(\omega) = \sum_{k\in\mathbb{Z}}\Bigg[\sum_{n=-N}^{N}\left[f_k[n]\left(e^{(2\pi int/T_c)}\right)\widehat{(}\omega)\right] * \left[\chi_{[kT_c,(k+1)T_c]}(t)\right]\widehat{(}\omega)\Bigg]$$

Performing the indicated transforms using the definition results in

$$\widehat{f_{\mathcal{P}}}(\omega) = \sum_{k\in\mathbb{Z}}\Bigg[\sum_{n=-N}^{N} f_k[n]\left(\delta(\omega-\frac{n}{T_c})\right) * e^{(2\pi i(k-\frac{1}{2})T_c\omega)}\frac{\sin(\pi T_c\omega)}{\pi\omega}\Bigg]$$

It is important to note that $f\cdot\chi_{[kT_c,(k+1)T_c]}$ is no longer band-limited, but it does decay at a rate less than or equal to $\frac{1}{\omega}$ in frequency. Using the relationship between translation and modulation, we get the following.

Theorem : [The Fourier Transform of the Projection Formula] Let $f,\ \widehat{f}\in L^2(\mathbb{R})$ and $f\in PW_\Omega$, i.e. $\mathrm{supp}(\widehat{f})\subset[-\Omega,\Omega]$. Let T_c be a fixed block of time. Then, for $N=\lceil T_c\cdot\Omega\rceil$,

$$\widehat{f_{\mathcal{P}}}(\omega) = \sum_{k=-\infty}^{\infty}\Bigg[\sum_{n=-N}^{N} f_k[n]e^{(2\pi i(k-\frac{1}{2})T_c(\omega-\frac{n}{T_c})} \left(\frac{\sin(\pi(\frac{\omega T_c}{2}-\frac{n}{2}))}{\pi(\omega-\frac{n}{T_c})}\right)\Bigg] \tag{4}$$

The system using overlapping Trap functions has the advantage of $1/\omega^2$ decay in frequency. Let $\beta_L = \sqrt{T_c/(4r)+r}$, $\alpha_L = T_c/(4r)+r/2$, $\beta_S = \sqrt{T_c/(4r)+r-1}$, $\alpha_S = T_c/(4r)-r/2$. The Fourier transform of Trap is

$$\left[(\beta_L)\frac{\sin(2\pi\alpha_L(\omega)}{\pi(\omega)}\right]^2 - \left[(\beta_S)\frac{\sin(2\pi\alpha_S(\omega)}{\pi(\omega)}\right]^2.$$

This replaces the sinc term in the equation above. The Fourier coefficients are also different, and are computed in the same method as the de la Vallée-Poussin kernel used in Fourier series.

In the formula for the Projection Method, there is a reliance on a number N, representative of the number of Fourier series components. In order to ensure maximum utility from the formula, the difference between the infinitely summed series and the truncated must be made a minimum. To do this, the mean square error must be calculated. We compute this as a truncation error on the number of Fourier coefficients used to represent a given block f_k. For a fixed N, the mean square error is

$$e_N^2 = \|f_k - f_{k,N}\|_2^2 = \|\widehat{f_k} - \widehat{f_{k,N}}\|_2^2.$$

Computing and then simplifying gives

$$e_N^2 = \frac{1}{T_c}\int_{kT_C}^{(k+1)T_c} |f_k{}^\circ(t) - \sum_{|n|\le N} f_k[n]e^{(2\pi itn/T_c)}|^2dt$$

$$= \frac{1}{T_c}\int_{kT_C}^{(k+1)T_c} |\sum_{|n|>N} f_k[n]e^{(2\pi itn/T_c)}|^2dt\,.$$

Applying the triangle inequality to the right side and then exploiting the fact that $e^{(2\pi itn/T_c)}$ is an orthonormal system, thus $|e^{(2\pi itn/T_c)}|=1$, we arrive at the following:

$$e_N^2 = \frac{1}{T_c}\int_{kT_C}^{(k+1)T_c} |\sum_{|n|>N} f_k[n]e^{(2\pi itn/T_c)}|^2dt \tag{5}$$

$$\le \sum_{|n|>N}|f_k[n]|^2\cdot\frac{1}{T_c}\int_{kT_c}^{(k+1)T_c}1^2dt = \sum_{|n|>N}|f_k[n]|^2$$

This demonstrates that the value of N has to be chosen carefully. This truncation error perpetuates over all the blocks.

The Projection Method experiences error due to truncation in two separate categories: time and frequency. The error in frequency is a function of the errors on each block due to the choice of N. By duality, this gives us errors in time. We can also get an error in time by loss of a given block or blocks of information. This is easier to compute. Given any lost or partially transmitted block $f_{k,L}$, error is simply

$$\|f_k - f_{k,L}\|_2\,.$$

Error over the entire signal is computed by simply adding up the blocks. Cell phone users are used to lost information blocks, which gives rise to the following frequently used phrase – "Can you hear me now?"

4. Binary Signals

The Walsh functions $\{\omega_n\}$ form an orthonormal basis for $L^2[0,1]$. The basis functions have the range $\{1,-1\}$, with values determined by a dyadic decomposition of the interval. The Walsh functions are of modulus 1 everywhere. The functions are give by the rows of the unnormalized Hadamard matrices, which are generated recursively by

$$H(2) = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

$$H(2^{(k+1)}) = H(2)\otimes H(2^k) = \begin{bmatrix} H(2^k) & H(2^k) \\ H(2^k) & -H(2^k) \end{bmatrix}.$$

We point out that although the rows of the Hadamard matrices give the Walsh functions, the elements have to be reordered into *sequency* order. Walsh arranged the components in ascending order of zero crossings (see [1]). The Walsh functions can also be interpreted as the characters of the group G of sequences over $\mathbb{Z}_2$, i.e., $G=(\mathbb{Z}_2)^{\mathbb{N}}$. The Walsh basis is a well-developed system for the study of a wide variety of signals, including binary. The Projection Method works with the Walsh system to create a wavelet-like system to do signal analysis.

First assume that the time domain is covered by a uniform block tiling $\chi_{[kT_c,(k+1)T_c]}(t)$. Translate and scale the function on this kth interval back to $[0,1]$ by a linear mapping. Denote the resultant mapping as f_k, which is an element of $L^2[0,1]$. Given that $f\in PW(\Omega)$, there exists an $N>0$ ($N=N(\Omega)$) such that $\langle f_k,\omega_n\rangle = 0$ for all $n>N$. The decomposition of f_k into Walsh basis elements is

$$\sum_{n=0}^{N}\langle f_k,\omega_n\rangle\,\omega_n\,.$$

Translating and summing up gives the Projection representation $f_{\mathcal{P}}$

$$f_{\mathcal{P}}(t) = \sum_{k\in\mathbb{Z}} \left[\sum_{n=0}^{N} \langle f_k, \omega_n \rangle \, \omega_n \right] \chi_{[kT_c,(k+1)T_c]}(t). \quad (6)$$

Next assume that the time domain is covered by a uniform overlapping trapezoidal tiling $\mathrm{Trap}(t - k(T_c/2))$. Note that the construction of the trapezoidal system results in the loss of no signal data, for just as a given block is ramping down, the subsequent block is ramping up at exactly the same rate. Again translate and scale the function on this kth interval back to $[0, 1]$ by a linear mapping. Denote the resultant mapping as f_{k_T}. The resultant function is an element of $L^2[0, 1]$. Given that $f \in PW(\Omega)$, there exists an $M > 0$ ($M = M(\Omega)$) such that $\langle f_{k_T}, \omega_n \rangle = 0$ for all $n > M$. The decomposition of f_{k_T} into Walsh basis elements is

$$\sum_{n=0}^{M} \langle f_k, \omega_n \rangle \, \omega_n \, .$$

Translating and summing up gives the Projection representation $f_{\mathcal{P}_T}$

$$f_{\mathcal{P}_T}(t) = \sum_{k\in\mathbb{Z}} \left[\sum_{n=0}^{N} \langle f_{k_T}, \omega_n \rangle \, \omega_n \right] \mathrm{Trap}(t - k(T_c/2)). \quad (7)$$

5. Conclusions

The Projection Method gives a method for analog-to-digital encoding which is an alternative to Shannon Sampling. Projection gives a procedure for the sampling of a signal of variable or ultra-wide bandwidth Ω by varying the time blocks T_c. If f is Ω band-limited, we can estimate the value of n for which the Fourier coefficients $f_k[n]$ of a given time block are non-zero. At minimum, $f_k[n]$ is non-zero if $\frac{n}{T_c} \leq \Omega$, or equivalently, $n \leq T_c \cdot \Omega$. If $N = \lceil T_c \cdot \Omega \rceil$, then, $f(t) \approx f_{\mathcal{P}}(t)$, where

$$f_{\mathcal{P}}(t) = \sum_{k\in\mathbb{Z}} \left[\sum_{n=-N}^{N} f_k[n] e^{(2\pi i n t/T_c)} \right] \chi_{[kT_c,(k+1)T_c]}(t).$$

For fixed N, if greater bandwidth Ω is need, choose shorter time blocks T_c. The price paid for this flexibility is in signal error, which has been computed above. The Projection Method can also adapt to changes in the signal, e.g., $f(t)$ has a band-limit $\Omega(t)$ which changes with time. This change effects the time blocking $\tau_c(t)$ and the number of basis elements $N(t)$. During a given $\tau_c(t)$, let $\overline{\Omega}(t) = \sup\{\Omega(t) : t \in \tau_c(t)\}$. For a signal f that is $\Omega(t)$ band-limited, we can estimate the value of n for which $f_k[n]$ is non-zero. At minimum, $f_k[n]$ is non-zero if

$$\frac{n}{\tau_c(t)} \leq \overline{\Omega}(t) \, , \text{ or equivalently, } n \leq \tau_c(t) \cdot \overline{\Omega}(t) \, .$$

We let

$$N(t) = \lceil \tau_c(t) \cdot \overline{\Omega}(t) \rceil \, ,$$

and have

$$f_{\mathcal{P}}(t) = \sum_{k\in\mathbb{Z}} \left[\sum_{n=-N(t)}^{N(t)} f_k[n] e^{(2\pi i n t/\tau_c)} \right] \chi_{[k\tau_c,(k+1)\tau_c]}(t).$$

This adaptable time segmentation makes the analysis more complicated, but is at the heart of the advantage the Projection Method has over conventional methods. Subsequent work on this method will focus on minimizing error, creating systems based on the Projection Method tailored to different types of signals and optimizing signal reconstruction in a noisy environment.

References:

[1] Beauchamp, K. G., *Applications of Walsh and Related Functions*, Academic Press, London, 1984.

[2] Benedetto, J. J., *Harmonic Analysis and Applications*, CRC Press, Boca Raton, FL, 1997.

[3] Casey, S. D., "Sampling and reconstruction on unions of non-commensurate lattices via complex interpolation theory," *1999 International Workshop on Sampling Theory and Applications*, 48–53, 1999.

[4] Casey, S. D., and Sadler, B. M., "New directions in sampling and multi-rate A-D conversion via number theoretic methods," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2000)*, 3, 1417–1420, 2000.

[5] Casey, S. D., and Walnut, D. F., "Residue and sampling techniques in deconvolution," Chapter 9 in *Modern Sampling Theory: Mathematics and Applications*, Birkhauser Research Monographs, ed. by P. Ferreira and J. Benedetto, 193–217, Birkhauser, Boston, 2001.

[6] Casey, S. D., "Two Problems from Industry and Their Solutions via Harmonic and Complex Analysis, to appear in *The Journal of Applied Functional Analysis*, 31 pp., 2009.

[7] Hoyos, S., and Sadler, B. M. "Ultra wideband analog-to-digital conversion via signal expansion," *IEEE Transactions on Vehicular Technology,* Invited Special Section on UWB Wireless Communications, vol. 54, no. 5, pp. 1609–1622, September 2005.

[8] Hoyos, S., Sadler, B. .M., and Arce, G., "Broadband multicarrier communication receiver based on analog to digital conversion in the frequency domain," *IEEE Transactions on Wireless Communications,* vol. 5, no. 3, pp. 652–661, March 2006.

[9] Hoyos, S., and Sadler, B. .M. "Frequency domain implementation of the transmitted-reference ultra-wideband receiver," *IEEE Transactions on Microwave Theory and Techniques,* Special Issue on Ultra-Wideband, vol. 54, no. 4, Part II, pp. 1745–1753, April 2006.

[10] Hoyos, S., and Sadler, B. M. "UWB mixed-signal transform-domain direct-sequence receiver," *IEEE Transactions on Wireless Communications,* vol. 6, no. 8, pp. 3038-3046, August 2007.

Non-Uniform Sampling Methods for MRI

Steven Troxler

(1) Arizona State University, Tempe AZ 85287-1804 USA.
Steven.Troxler@asu.edu

1. Introduction

Simple Cartesian scans, which collect Fourier transform data on a uniformly-spaced grid in the frequency domain, are by far the most common in MRI. But non-Cartesian trajectories such as spirals and radial scans have become popular for their speed and for other benefits, like making motion-correction easier [12]. A major problem in such scans, however, is reconstructing from nonuniform data, which cannot be performed by a standard fast Fourier transform (FFT) as in the Cartesian case.

Here, we briefly describe the most common reconstruction methods and the non-uniform fast Fourier transform (NFFT) needed to complete the computations quickly. We then give an overview of several current methods for choosing a density compensation function (DCF) and suggest some possible improvements.

2. Reconstruction Methods

The most common method for nonuniform reconstruction in MRI is the Riemann approach, which approximates the integral difining the inverse (continuous) Fourier transform using a Riemann sum

$$f_{\boldsymbol{w}}(\boldsymbol{x}) = \sum_{j=1}^{J} w_j \hat{f}(\boldsymbol{\xi}_j) e^{2\pi i \boldsymbol{\xi}_j \cdot \boldsymbol{x}}, \tag{1}$$

where $\boldsymbol{x} \in \mathbb{Z}_N^d$ are the pixel locations and $\boldsymbol{\xi}_j$, $j = 1, ..., J$, are the frequency locations at which we measure the Fourier transform (we assume $J \geq N^d$). As the subscript $\boldsymbol{w}$ suggests, this approach requires finding appropriate weights w_j for each sample point in the reconstruction, a major theoretical problem. An alternative method, called implicit discretization (ID), assumes that the image itself is a sum of evenly spaced delta impulses at the pixel points of the final image, so that its Fourier transform is a finite-dimensional, harmonic trigonometric polynomial. We can then find a least-squares solution to the resulting system of equations

$$\hat{f}(\boldsymbol{\xi}_j) = \sum_{\boldsymbol{x} \in \mathbb{Z}_N^d} f(\boldsymbol{x}) e^{-2\pi i \boldsymbol{x} \cdot \boldsymbol{\xi}_j} \tag{2}$$

This model, which is known to have negligible error (the model error is the Gibb's error that would appear in a Cartesian reconstruction), has the important advantage of not depending on our arbitrary choice of weights.

These two approaches can be described in terms of matrix algebra as follows: Let G be a $J \times N^d$ matrix given by

$$\boldsymbol{G}_{j,\boldsymbol{x}} = e^{-2\pi i \boldsymbol{\xi}_j \cdot \boldsymbol{x}}.$$

Then we see immediately that

$$\boldsymbol{f}_{\boldsymbol{w}} = \boldsymbol{G}^* \boldsymbol{W} \tilde{\boldsymbol{f}}, \tag{3}$$

where $\boldsymbol{f}_{\boldsymbol{w}}$ is the $N^d \times 1$ vector, indexed by $\mathbb{Z}_N^d$, whose $\boldsymbol{x}$th entry is $f_{\boldsymbol{w}}(\boldsymbol{x})$, $\tilde{\boldsymbol{f}}$ is the $J \times 1$ vector of measurements whose jth entry is $\hat{f}(\boldsymbol{\xi}_j)$, and $\boldsymbol{W}$ is the $N^d \times N^d$ diagonal matrix with diagonal equal to $\boldsymbol{w}$. Once we have $\boldsymbol{w}$, whose determination is the main problem of interest, the remaining issue is one of computational complexity, since $\boldsymbol{G}^*$ is a very large unstructured matrix.

Fortunately, there is a fast method for computing products called the nonuniform fast Fourier transform (NFFT), based on the approximate factorization

$$\boldsymbol{G} \approx \boldsymbol{C}_\phi \boldsymbol{F} \boldsymbol{D}_\phi, \tag{4}$$

where $\boldsymbol{C}_\phi$ is a sparse, banded $N^d \times J$ matrix of convolution interpolation coefficients which depends on our choice of convolution kernel ϕ, $\boldsymbol{F}$ is the uniform $M^d \times M^d$ DFT matrix for some $M > N$, products of which are rapidly computed via the FFT, and $\boldsymbol{D}_\phi$ is an $M^d \times N^d$ modified diagonal deconvolution matrix, also depending on ϕ, whose extra rows are zero. Since it is easy to compute products with all three factors, this algorithm can be used to quickly approximate matrix products involving either $\boldsymbol{G}$ or $\boldsymbol{G}^*$. The theory of the NFFT, as applied to MRI, was first laid out in [11] and [8]. Later, [4] found bounds on the errors for Gaussian interpolation, and [23] and [5] gave general estimates and gave sharper bounds for Gaussian kernels. The most complete discussion of NFFT theory is given in [16], while [15] presents many of the proofs. Practical considerations like computational load and numerical stability were addressed in [3] and [17], while [1] and [6] presented two methods of efficient interpolation using Kaiser-Bessel and Gaussian kernels.

In matrix form, the ID problem attempts to find a least-squares solution to the problem

$$\tilde{\boldsymbol{f}} = \boldsymbol{G} \boldsymbol{f}.$$

The ordinary least squares solution $\boldsymbol{f}_{OLS}$ satisfies the normal equation

$$\boldsymbol{G}^*\boldsymbol{G}\boldsymbol{f}_{OLS} = \boldsymbol{G}^*\hat{\boldsymbol{f}}. \tag{5}$$

Although the matrix $\boldsymbol{G}^*\boldsymbol{G}$ is far too large to invert, it is symmetric, so we may use iterative methods like conjugate gradients to find the solution. The resulting solution typically has excellent quality, but convergence is often slow, making ordinary least squares expensive.
Conjugate gradients converges fastest when $\boldsymbol{G}^*\boldsymbol{G}$ is close to the identity, which is unfortunately rarely the case unless the sampling density is reasonably close to unity. In order to improve the convergence of conjugate gradients, we introduce the weighted least squares problem, which finds the least squares solution to

$$\boldsymbol{W}^{1/2}\tilde{\boldsymbol{f}} = \boldsymbol{W}^{1/2}\boldsymbol{G}\boldsymbol{f}$$

by solving the normal equations

$$\boldsymbol{G}^*\boldsymbol{W}bvecG\boldsymbol{f}_{WLS} = \boldsymbol{G}^*\boldsymbol{W}\hat{\boldsymbol{f}},$$

where $\boldsymbol{W}$ is the modified diagonal density compensation matrix used for the Riemann method. We expect an improvement in convergence because we know that the Riemann method gives much better results with $\boldsymbol{W}$ than without, which means $\boldsymbol{G}^*\boldsymbol{W}\boldsymbol{G}$ approximate the identity much better than $\boldsymbol{G}^*\boldsymbol{G}$. From a signal processing perspective, this has the additional benefit that we weight errors heavier at highly isolated observations of the Fourier transform, which heuristically contain more information about the objective function than less isolated observations.
For either method, then, determining an appropriate value of $\boldsymbol{w}$ is important. It is more essential in the Riemann approach, where a poor choice of $\boldsymbol{w}$ will lead to useless results. The ID method is known to converge quite well after only a few iterations, even when a very rough approximation to $\boldsymbol{w}$ is used, but the better $\boldsymbol{w}$, the fewer iterations are required. It is worth noting that the first iteration, which always moves in the direction of the residual, is actually just a rescaling of the Riemann solution.

3. Determination of an optimal DCF

3.1 Algebraic and Analytic Approaches

Since the equation

$$\tilde{\boldsymbol{f}} = \boldsymbol{G}\boldsymbol{f},$$

used directly in the CG reconstruction, provides an accurate mathematical model for the measurements which does not depend on the choice of a sampling density $\boldsymbol{w}$, the clearest method of evaluating a DCF $\boldsymbol{w}$ is to require that

$$\tilde{\boldsymbol{f}} \approx \boldsymbol{G}\boldsymbol{f}_{\boldsymbol{w}},$$

where

$$\boldsymbol{f}_{\boldsymbol{w}} = \boldsymbol{G}^*\boldsymbol{W}\tilde{\boldsymbol{f}}.$$

This is the same as requiring that

$$\boldsymbol{G}^*\boldsymbol{W}$$

approximate the pseudoinverse $(\boldsymbol{G}^*\boldsymbol{G})^{-1}\boldsymbol{G}^*$ of $\boldsymbol{G}$.

The weighted conjugate gradient method described at the end of the previous section, whose first iteration performs best when the matrix is as close to the identity as possible, leads to a similar but slightly simpler condition, that

$$\boldsymbol{G}^*\boldsymbol{W}\boldsymbol{G} \approx \boldsymbol{I},$$

in the sense that the eigenvalues of $\boldsymbol{G}^*\boldsymbol{W}\boldsymbol{G}$ be as closely clustered as possible. Several techniques have been proposed to use these conditions to find an algebraically ideal DCF via use of a singular value decomposition or some similar approach [22], [20]. These methods, however, tend to have high computational complexity. This is a problem if the same trajectory is not always used, as is the case in many MRI applications in which iterative reconstruction is used to compensate for field inhomogeneities and other measurement imperfections. Moreover, although such algebraic methods generally give workable results, other methods which take analytic considerations into account often perform better empirically. Possible reasons why the theoretically optimal algebraic solutions fail to give the best results include numerical instability and ill-conditioning. In some cases, the algebraic approaches even result in DCF's with negative weights at some points. This contradicts our intuition, and empirical studies indicate that such DCF's tend to perform relatively poorly.
The simplest analytic approaches to determining $\boldsymbol{w}$ are based on the fact that the goal of the Riemann method is to approximate a Riemann sum. For radial and analytic spiral trajectories, which may be smoothly parameterized, methods have been proposed which use the Jacobian of a change-of-coordinates [10], [7]. These techniques give very good results for certain spirals, although for radial trajectories they tend to underweight points near the center. An alternative analytic method, which works for arbitrary nonuniform sampling schemes, is to construct a Voronoi diagram, which partitions the sampled part of frequency space into polygons about each sample point, and weight the samples according to the area or volume of those polygons [19]. This typically results in a good image for radial trajectories. With other trajectories, the results are generally inferior to alternative point-spread-function methods, although it was demonstrated in [9] that performing a few iterations of the weighted conjugate gradient method using Voronoi weights produces an excellent image.

3.2 The Point Spread Function

Most of the best-performing methods for determining the DCF when the trajectory is anything other than an analytic spiral are based on analysis of the *point-spread-function* (PSF). The PSF is defined as the inverse Fourier transform $\check{w}$ of the DCF, where we view the DCF as a distribution on $\mathbb{R}^d$ defined by $w := \sum_j w_j \delta_{\boldsymbol{\xi}_j}$. The PSF $\check{w}$ is then given by

$$\check{w}(\boldsymbol{x}) = \sum_{j=1}^{J} w_j e^{2\pi i \boldsymbol{x}\cdot\boldsymbol{\xi}_j}. \tag{6}$$

This is what the algorithm would produce if the true object were a delta impulse located at zero. The observed data would be a vector of all ones, so the reconstruction would

be the result of applying G^* to $\boldsymbol{w}$ itself, i.e., the function defined by (6).

If f is a more general object, it follows from the convolution theorem (for distributions) that the reconstructed function $f_{\boldsymbol{w}}$ will be equal to the convolution $f * \check{w}$ of the actual object f with the PSF. The more closely the PSF resembles a delta impulse, the better the reconstruction.

It is important to note that, since the PSF is a (nonharmonic) trigonometric polynomial, it will not decay at infinity. Clearly, then, the best that we can hope for is that $\check{w}$ will resemble a delta impulse in some compact neighborhood of the origin. Recall that, by accepting the ID model as having negligible error, we are assuming that f is a finite-dimensional vector defined on $\mathbb{Z}_N^d$ which we associate with a distribution supported on $\mathbb{Z}_N^d$ for notational convenience when dealing with convolutions. Since the terms $\check{w}(\boldsymbol{z})f(\boldsymbol{x}-\boldsymbol{z})$ defining

$$f_{\boldsymbol{w}}(\boldsymbol{x}) = \check{w} * f(\boldsymbol{x}) = \sum_{\boldsymbol{z}\in\mathbb{Z}^d} \check{w}(\boldsymbol{z}) f(\boldsymbol{x}-\boldsymbol{z}) \qquad (7)$$

are nonzero only if $(\boldsymbol{x}-\boldsymbol{z}) \in \mathbb{Z}_N^d$, and we only want to find the reconstruction $f_{\boldsymbol{w}}(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathbb{Z}_N^d$, we conclude that the only values of $\boldsymbol{z}$ for which $\check{w}(z)$ matter are $\boldsymbol{z} \in \mathbb{Z}_{2N}^d$. It is also worth noting that not all points $\boldsymbol{z} \in \mathbb{Z}_{2N}^d$ appear equally often in the convolution defining $\boldsymbol{f}_W$. The origin will appear in one term of every sum, whereas values of $\boldsymbol{z}$ near the edge of $\mathbb{Z}_{2N}^d$ will appear only occasionally.

For notational convenience, let A be the field of view $[-N/2, N/2]^d$ and let B be the region of optimization $[-N, N]^d$. PSF optimization techniques find some computational way of minimizing the error

$$E = \check{w} - \delta$$

over this region of optimization B.

By carefully looking at (7), we see that the frequency with which a PSF error at $\boldsymbol{x}$ actually occurs in the final image is proportional to $p = \chi A * \chi A$. Since errors are unavoidable and we would like to minimize the important errors, we introduce the weighted error, given by

$$E = p\check{w} - \delta, \qquad (8)$$

where p is called the *error profile*. This error can be expressed in the Fourier domain as

$$\hat{E} = \hat{p} * w - 1 = \hat{\chi}_A^2 * w - 1. \qquad (9)$$

Our goal is to minimize these errors, thereby minimizing the error in the final reconstruction $f_{\boldsymbol{w}} = \check{w} * f$.

Although this optimal kernel p was suggest only recently in [13], convolution techniques for minimizing the Fourier domain PSF error $\hat{E}$ have been used for some time. In one of the early gridding papers, Jackson et. al. proposed taking w to be equal to

$$w_1 = \frac{w_0}{\phi * w_0}, \qquad (10)$$

where w_0 is a DCF of unity (in distributional form) and ϕ is the gridding kernel [8]. This method predates PSF techniques, and was instead motivated by the intuitive idea that $\phi * w_0$ would give a reasonable, estimate of the sampling density. Later researchers noted, however, that if we ϕ with $\hat{p}$, we would expect this ratio correction to make $w_1 * \hat{p}$ closer to unity than $w_0 * \hat{p}$ regardless of the initial density w_0 [14]. An iterative technique, based on this observation, starts with a constant DCF w_0 and takes

$$w_{i+1} = \frac{w_i}{\hat{p} * w_i}. \qquad (11)$$

Since $\hat{p}$ can be effectivly truncated, each iteration can be computed quickly, particularly if an efficient sorting algorithm is used to avoid time-consuming searches for the nonzero terms $\hat{p}(\boldsymbol{\xi_k} - \boldsymbol{\xi}_j)w(\boldsymbol{\xi}_j)$ in the convolution [13].

Another iterative algorithm, aimed at the same goal of achieving $\hat{E} = 1$, uses an additive correction instead of a ratio-based correction, taking

$$w_i = w_{i-1} + \sigma(1 - \hat{p} * w_{i-1}),$$

where $\sigma \in (0, 1)$ is a parameter controlling convergence [18]. Taking σ close to 1 may result in the fastest convergence, but could also lead to instability and a failure to converge.

The advantage of these iterative techniques is that they are conceptually simple, computationally fast, and empirically give results as good as any current methods when the correct error profile p is used and the number of iterations is determined experimentally. A disadvantage is that, although they work conceptually and empirically, there is no theoretical basis for claiming that they converge to the optimal solution, and, in fact, experimental evidence indicates that the mean square error in $f_{\boldsymbol{w}}$ can actually rises if the algorithm is allowed to run too long. This may be due to numerical instability, or to a failure of the mathematical algorithm itself to technically converge.

An algebraic method of optimizing the PSF, which has more theoretical grounding than convolution-based methods, attempts to directly solve the inverse problem

$$\boldsymbol{G}\boldsymbol{G}^*\boldsymbol{w} = \boldsymbol{u},$$

where $\boldsymbol{u}$ is a vector of all ones. The direct solution to this problem via conjugate gradients using the NFFT was proposed in [21], but as with the algebraic solutions for $\boldsymbol{w}$ based on the least-squares method, this can result in a $\boldsymbol{w}$ with wide variations and sometimes even negative entries, which does not match our expectation for a density and empirically gives inferior results. A regularization of this method was proposed in [2] which instead attempts to solve

$$(\boldsymbol{G}\boldsymbol{G}^* + \sigma^2\boldsymbol{I})\boldsymbol{w} = \boldsymbol{u} + \sigma^2\boldsymbol{w}_1,$$

where $\boldsymbol{w}_1$ is an initial nonnegative and smoothly varying estimate of the density, say, Jackson's weight (10), or, more optimally, the result of one or two iterations of (11). This second approach ensures that the solution behaves as we would expect a DCF to behave, and empirically gives better results than the unregularized method. The algorithm given in [2] also incorporates Jacobi preconditioning to speed convergence of the conjugate gradient iterations.

Knowing that Pipe and Johnson's error profile p provides an optimal weight on errors in the point-spread function,

it might be preferable to modify the approach in [2] in two ways. The first is that, since we need to minimize PSF errors over twice the support of f, we replace the NDFT matrix $\boldsymbol{G}$ with $\boldsymbol{G}_1$, where the uniform grid has twice the radius of that used by $\boldsymbol{G}$. This avoids the risk that we might ignore PSF errors which, according to the convolution defining $\boldsymbol{f_w}$, appear in the Riemann reconstruction. The second is replacing $\boldsymbol{G}_1\boldsymbol{G}_1^*$, which treats all PSF errors as equally important, with $\boldsymbol{G}_1\boldsymbol{P}\boldsymbol{G}_1^*$, where $\boldsymbol{P}$ contains the values of the optimal error profile p. To the author's knowledge, this has never been tried, but in light of experiments reported by [13] indicating that the approaches taken in [2] and [13] both yield the some of the best results of methods proposed to date for arbitrary trajectories, combining their methods might produce the best results seen yet.

4. Acknowledgments

This work was part of an undergraduate research project under the supervision of Dr. Svetlana Roudenko. The author would also like to thank Ken Johnson and Dr. Jim Pipe for providing graphical illustrations and for helpful discussions of this content, as well as Dr. Doug Cochran for his assistance and input. The project was partially supported by NSF-DUE # 0633033 and NSF-DMS # 0652853.

References:

[1] Philip J. Beatty, Dwight G. Nishimura, and John M. Pauly. Rapid gridding reconstruction with a minimal oversampling ratio. *IEEE Trans. Med. Imag.*, 24(6):799–808, June 2005.

[2] Mark Bydder, Alexey A. Samsonov, and Jiang Du. Evaluation of optimal density weighting for regridding. *Magnetic Resonance Imaging*, 25:695–702, 2007.

[3] S. Dunis and D. Potts. Time and memory requirements of the nonequispaced fft. *Sampling Theory in Signal and Image Processing*, 7:77–100, 2008.

[4] A. Dutt and V. Rokhlin. Fast fourier transforms for nonequispaced data. *SIAM Journal of Scientific Computing*, 14(6):1368–1393, 1993.

[5] B. Elbel and G. Steidl. Fast fourier transform for nonequispaced data. In C. K. Chui and L. L. Schumaker, editors, *Approximation Theory IX*. Vanderbilt University Press, Nashville, 1998.

[6] Leslie Greengard and June-Yub Lee. Accelerating the nonuniform fast fourier transform. *SIAM Review*, 46(3):443–454, 2004.

[7] R.D Hodge, R.K.S. Kwan, and G.B. Pike. Density compensation functions for spiral MRI. *Magnetic Resonance in Medicine*, 38:117–128, 1997.

[8] J. Jackson, C. Meyer, D. Nishimura, and A. Macovski. Selection of a convolution function for fourier enversion using gridding. *IEEE Trans. Med. Imag.*, 10(3):473–478, Sep. 1991.

[9] Tobias Knopp, Stefan Kunis, and Daniel Potts. A note on the iterative mri reconstruction from nonuniform k-space data. *International Journal of Biomedical Imaging*, 2007.

[10] C. Meyer, B. S. Hu, D. Nishimura, and A. Macovski. Fast spiral coronary artery imaging. *Magnetic Resonance in Medicine*, 28:202–213, 1992.

[11] J. O'sullivan. A fast sinc function gridding algorithm for fourier inversion in computerized tomography. *IEEE Transactions on Medical Imaging*, MI-4:200–207, 1985.

[12] James G. Pipe.

[13] J.G. Pipe and Kenneth Johnson. Convolution kernel design and efficient algorithm for sampling density correction. *Preprint*, 2008.

[14] J.G. Pipe and P. Menon. Sampling density compensation in MRI: Rationale and an iterative numerical solution. *Magnetic Resonance in Medicine*, 41:799–808, June 2005.

[15] D. Potts. *Schnelle Fourier-Transformationen für nichtäquidistante Daten und Anwendungen*. Habilitation, Universität zu Lübeck, 2003.

[16] D. Potts, G. Steidl, and M. Tasche. Fast fourier transforms for nonequispaced data: A tutorial. In J. J. Benedetto and P. J. S. G. Ferreira, editors, *Modern Sampling Theory: Mathematics and Applications*, pages 247–270. Birkhäuser, Boston, 2001.

[17] D. Potts and M Tasche. Numerical stability of nonequispaced fast fourier transforms. *Journal of Computational Applied Mathematics*, 222:655–674, 2008.

[18] Y. Qian, J. Lin, and D. Jin. Reconstruction of mr images from data acquired on an arbitrary k-space trajectory using the same-image weight. *Magnetic Resonance in Medicine*, 48:306–311, 2002.

[19] V. Rasche, R. Proksa, R. Sinkus, P. Bornert, and H. Eggers. Resampling of data between arbitrary grids using convolution interpolation. *IEEE Trans. Med. Imag.*, 18:427–434, 1999.

[20] D. Rosenfeld. An optimal and efficient new gridding algorithm using singular value decomposition. *Magnetic Resonance in Medicine*, 40:14–23, 1998.

[21] A.A. Samsonov, E.G. Kholmovski, and C.R. Johnson. Determination of the sampling density compensation function using a point spread function modeling approach and gridding approximation. volume 11, 2003.

[22] Hossein Sedarat and Dwight G. Nishimura. On the optimality of the gridding reconstruction algorithm. *IEEE Transactions on Medical Imaging*, 19(4):306–317, 2000.

[23] G. Steidl. A note on fast fourier transforms for nonequispaced grids. *Advanced Computational Mathematics*, 9:337–353, 1998.

On approximation properties of sampling operators defined by dilated kernels

Andi Kivinukk [(1)] and Gert Tamberg [(2)]

(1) Dept. of Mathematics, Tallinn University, Narva Road 25, 10120 Tallinn, Estonia.
(2) Dept. of Mathematics, Tallinn University of Technology, Ehitajate tee 5 19086 Tallinn, Estonia.
andik@tlu.ee, gert.tamberg@mail.ee

Abstract:

In this paper we consider some generalized Shannon sampling operators, which are defined by band-limited kernels. In particular, we use dilated versions of some previously known kernels. We give also some examples of using sampling operators with dilated kernels in imaging applications.

1. Introduction

For the uniformly continuous and bounded functions $f \in C(\mathbb{R})$ the generalized sampling series with a kernel function $s \in L^1(\mathbb{R})$ are given by ($t \in \mathbb{R}$; $W > 0$)

$$(S_W f)(t) := \sum_{k=-\infty}^{\infty} f(\frac{k}{W}) s(Wt - k) \qquad (1)$$

where

$$\sum_{k=-\infty}^{\infty} s(u-k) = 1, \qquad (2)$$

and their operator norms are

$$\|S_W\| = \sum_{k=-\infty}^{\infty} |s(u-k)| < \infty \quad (u \in \mathbb{R}).$$

If the kernel function is $s(t) = \operatorname{sinc}(t) := \frac{\sin \pi t}{\pi t}$, we get the classical (Whittaker-Kotel'nikov-)Shannon operator $S_W^{\operatorname{sinc}}$. The idea to replace the sinc kernel ($\operatorname{sinc}(\cdot) \notin L^1(\mathbb{R})$) by another kernel function $s \in L^1(\mathbb{R})$ appeared first in [15], where the case $s(t) = (\operatorname{sinc}(t))^2$ was considered. A systematic study of sampling operators (1) for arbitrary kernel functions s was initiated at RWTH Aachen by P. L. Butzer and his students since 1977 (see [3], [4], [14] and references cited there).

In this paper we consider the generalized sampling series with even band-limited kernels s, defined as the Fourier transform of an even window function $\lambda \in C_{[-1,1]}$, $\lambda(0) = 1$, $\lambda(u) = 0$ ($|u| \geqslant 1$) by the equality

$$s(t) := s(\lambda; t) := \int_0^1 \lambda(u) \cos(\pi t u)\, du = \sqrt{\frac{\pi}{2}} \lambda^\wedge(\pi t). \qquad (3)$$

These types of kernels arise in conjunction with window functions widely used in applications (e.g. [1], [2], [11], [16]), in Signal Analysis in particular. Many kernels can be defined by (3), e.g.

1) $\lambda(u) = 1$ defines the sinc function;

2) $\lambda(u) = 1 - u$ defines the Fejér kernel (cf. [15])

$$s_F(t) = \frac{1}{2} \operatorname{sinc}^2 \frac{t}{2} = O(|t|^{-2});$$

3) $\lambda_H(u) := \cos^2 \frac{\pi u}{2} = \frac{1}{2}(1 + \cos \pi u)$ defines the Hann kernel (see [7])

$$s_H(t) := \frac{1}{2} \frac{\operatorname{sinc} t}{1 - t^2} = O(|t|^{-3});$$

4) the general cosine window

$$\lambda_{C,\mathbf{b}}(u) := \sum_{j=0}^{m} b_j \cos j\pi u \qquad (4)$$

defines the Blackman-Harris kernel (see [9])

$$s_{C,\mathbf{b}}(t) := \frac{1}{2} \sum_{j=0}^{m} b_j \Big(\operatorname{sinc}(t-j) + \operatorname{sinc}(t+j) \Big) \qquad (5)$$

provided

$$\sum_{j=0}^{\lfloor \frac{m}{2} \rfloor} b_{2j} = \sum_{j=1}^{\lfloor \frac{m+1}{2} \rfloor} b_{2j-1} = \frac{1}{2}. \qquad (6)$$

From approximation theory point of view at least two problems for the generalized sampling operators $S_W : C(\mathbb{R}) \to C(\mathbb{R})$ have some interest:

1) to calculate the operator norms

$$\|S_W\| = \sup_{u \in \mathbb{R}} \sum_{k=-\infty}^{\infty} |s(u-k)|; \qquad (7)$$

2) to estimate the order of approximation

$$\|f - S_W f\|_C \leqslant M \omega_k(f, \frac{1}{W}) \qquad (8)$$

in terms of the k-th modulus of smoothness $\omega_k(f, \delta)$.

2. Interpolating generalized sampling operators with dilated kernels

Let us consider the dilated kernel $s_\alpha(t) = \alpha s(\alpha t)$. The Shannon operators with sinc kernel satisfy the interpolatory conditions

$$(S_W^{\operatorname{sinc}})(\frac{k}{W}) = f(\frac{k}{W}) \quad (k \in \mathbb{Z}). \qquad (9)$$

When we replace the sinc kernel with a band-limited one (3), we may lose the interpolatory property (9), but using the dilated kernel $\tilde{s}(t) = 2s(2t)$, we can recover the interpolatory property. If $s \in B^1_\pi$, then $s_\alpha \in B^1_{\alpha\,\pi}$, and the condition (2) is valid for $0 < \alpha \leqslant 2$, therefore we get the sampling operator $S_{W,\alpha} : C(\mathbb{R}) \to B^\infty_{\alpha\pi W} \subset C(\mathbb{R})$. Here B^p_σ stands for the Bernstein class consisting of those bounded functions $f \in L^p(\mathbb{R})$ $(1 \leqslant p \leqslant \infty)$ which can be extended to an entire function $f(z)$ $(z \in \mathbb{C})$ of exponential type σ.
Using the Nikolskii inequality [13], we get the bounds for the operator norm.

Theorem 1. *Let the operators $S_W : C(\mathbb{R}) \to B^\infty_{W\pi} \subset C(\mathbb{R})$, $S_{W,\alpha} : C(\mathbb{R}) \to B^\infty_{\alpha W\pi} \subset C(\mathbb{R})$ are defined by* (1) *with kernels s and s_α, respectively. Then*

$$\|s\|_1 \leqslant \|S_{W,\alpha}\| \leqslant (1+\alpha\pi)\|S_W\| \quad (0 < \alpha \leqslant 2).$$

The order of approximation by operators $S_{W,\alpha}$ we can estimate via modulus of smoothness $\omega_k(f,\sigma)$. Next theorem generalizes slightly the result in [10] (Th. 1.3).

Theorem 2. *Let $S_W : C(\mathbb{R}) \to C(\mathbb{R})$, $S_{W,\alpha} : C(\mathbb{R}) \to B^\infty_{\alpha W\pi} \subset C(\mathbb{R})$ be sampling operators defined by* (1) *with kernel functions $s \in B^1_\pi$, $s_\alpha \in B^1_{\alpha\,\pi}$, respectively.*
1) If $0 < \alpha \leqslant 1$, then there exist positive constants $C_{1,\alpha}$ and $C_{2,\alpha}$ such that

$$C_{1,\alpha}\|S_{\alpha W}f - f\|_C \leqslant \|S_{W,\alpha}f - f\|_C \leqslant C_{2,\alpha}\|S_{\alpha W}f - f\|_C.$$

2) Moreover, if $0 < \alpha < 2$, then

$$\|S_W f - f\|_C \leqslant M_k \omega_k(f, \frac{1}{W}), \tag{10}$$

implies

$$\|S_{W,\alpha}f - f\|_C \leqslant M_{k,\alpha}\omega_k(f, \frac{1}{W})$$

for some constant $M_{k,\alpha} > 0$.

Example. The Blackman-Harris sampling operator $C_{W,\mathbf{b}}$ is defined by the window function

$$\lambda_{C,\mathbf{b}} := \sum_{j=0}^{m} b_j \cos(\pi j u).$$

In [9] we proved that for some values of the parameters $\mathbf{b} = (b_0, b_1, \ldots, b_m) \in \mathbb{R}^{m+1}$ we can estimate the order of approximation by operators $C_{W,\mathbf{b}} : C(\mathbb{R}) \to B^\infty_{W\pi} \subset C(\mathbb{R})$ via the modulus of continuity $\omega_{2\ell}(f, \frac{1}{W})$ $(\ell \leqslant m)$. More precisely (see [9], Th. 3), let ℓ, $1 \leqslant \ell \leqslant m$, be fixed. If for every $k = 0, \ldots, \ell - 1$

$$\sum_{j=0}^{m} j^{2k} b_j = 0 \ \ (0^0 = 1), \tag{11}$$

then

$$\|f - C_{W,\mathbf{b}}f\|_C \leqslant M_{\mathbf{b},\ell}\omega_{2\ell}(f, \frac{1}{W}). \tag{12}$$

Now by Theorem 2 we obtain for the corresponding dilated sampling operator $C_{W,\mathbf{b};\alpha} : C(\mathbb{R}) \to B^\infty_{\alpha W\pi} \subset C(\mathbb{R})$ with $0 < \alpha < 2$ the estimate

$$\|f - C_{W,\mathbf{b};\alpha}f\|_C \leqslant M_{\mathbf{b},\ell,\alpha}\omega_{2\ell}(f, \frac{1}{W}). \tag{13}$$

The case $m = \ell = 1$ gives the Hann sampling operator $H_W : C(\mathbb{R}) \to C(\mathbb{R})$, which often has been used in practise. For the corresponding dilated operator $H_{W,\alpha} : C(\mathbb{R}) \to B^\infty_{\alpha W\pi} \subset C(\mathbb{R})$ for $0 < \alpha < 2$ we obtain

$$\|f - H_{W,\alpha}f\|_C \leqslant M_\alpha \omega_2(f, \frac{1}{W}). \tag{14}$$

See Figure 2 for corresponding kernels.
The next theorem gives hints how to construct the interpolating sampling series.

Theorem 3. *Let the sampling operator $\tilde{S}_W$ be defined by (1) using the kernel $\tilde{s}(t) := 2s(2t)$, where the kernel $s \in B^1_\pi \subset L^1(\mathbb{R})$ is generated by (3) with a window function λ. If*

$$\lambda(u) + \lambda(1-u) = 1 \ \ (u \in [0,1]) \tag{15}$$

then $\tilde{S}_W : C(\mathbb{R}) \to B^\infty_{2W\pi} \subset C(\mathbb{R})$ is an interpolating sampling operator.

Examples. For the Hann window function $\lambda_H(u)$ the condition (15) holds and we get the interpolating Hann sampling operator $\tilde{H}_W : C(\mathbb{R}) \to B^\infty_{2W\pi} \subset C(\mathbb{R})$. Taking $b_0 = 1/2$, $b_{2j} = 0 (j \in \mathbb{N})$ in (11) gives us the Blackman-Harris window function for which the condition (15) is fullfilled (see [10]).
In the case when $s \in B^1_{\beta\pi}$, $0 < \beta < 1$ and (15) holds for the corresponding window function we can prove the following theorem.

Theorem 4. *Let the sampling operator $\tilde{S}_W$ be defined by (1) using the kernel $\tilde{s}(t) := 2s(2t)$, where the kernel $s \in B^1_{\beta\pi} \subset L^1(\mathbb{R})$, $0 < \beta < 1$, is generated by (3) with a window function λ. If (15) is valid, then for every $k \in \mathbb{N}$ there exist a constant M_k such that*

$$\|\tilde{S}_W f - f\|_C \leqslant M_k \omega_k(f, \frac{1}{W}).$$

Example. So-called Lanczos n-kernels

$$\tilde{s}_{L,n}(t) := \operatorname{sinc}\frac{t}{n}\operatorname{sinc} t,$$

which has been often used in image processing. The Lanczos 3-kernel is especially popular in imaging ((see [16] and references cited there). They are defined by De la Vallée Poussin window function

$$\lambda_{L,n}(u) := \begin{cases} 1, & 0 \leqslant u \leqslant \frac{n-1}{2n}, \\ \frac{1}{2}(1 + n(1-2u)), & \frac{n-1}{2n} < u < \frac{n+1}{2n}, \\ 0, & u \geqslant \frac{n+1}{2n}. \end{cases}$$

If $n > 1$, then the De la Vallée Poussin window function $\lambda_{L,n}$ satisfies the conditions (15) and $\tilde{s}_{L,n} \in B^1_{(\frac{n+1}{2n})\pi}$, hence Theorem 4 is applicable. If $n = 1$, then we get the Fejér sampling operator (cf. [15]), for which we do not have even an estimate via the modulus of continuity ω_1.

3. Applications in 2D imaging

A natural application of sampling operators with dilated kernels is imaging. We can represent an discrete 2D image f as a continuous function using sampling series

$$(Sf)(x,y) := \sum_{j,k} f(j,k)s_1(x-j)s_2(y-k). \tag{16}$$

Figure 1: Original image, derivatives with Hann kernel $\tilde{s}_H(t) = 2s_H(2t)$ and $s_{H,1/4}(t) = \frac{1}{2}s_H(\frac{1}{4}t)$ ($\varphi = \frac{2\pi}{3}$).

Many image resizing (resampling) algorithms use such type of representation (see [16], [12], [6]). If the image data is exact, then we can take interpolating kernels s_1 and s_2, like interpolating Hann, Blackman-Harris or Lanczos, and enlarge (up-sample) image, having $(Sf)(j,k) = f(j,k)$. If we want to reduce the image size (down-sample) (magnification $\gamma < 1$) then, for eliminating artifacts, we can choose a dilated kernel s_α with in some sense optimal value of $\alpha = 2\gamma$ (see Figure 2). The artifacts in down-sampled images appear, because details that are resized to smaller than one pixel will be misrepresented by larger aliases (see [5], [6]). Depending on the choice of the parameter value α we have $S_{W,\alpha} : C(\mathbb{R}) \to B^\infty_{\alpha W\pi}$ i.e. a function belonging to a class for bandlimited functions, for which the Fourier' transform vanishes outside of the interval $[-\alpha W\pi, \alpha W\pi]$. This approach eliminates higher spatial frequencies, being equivalent to the use of low-pass filter. Also in the case, when the resolution of the optical system is less than the resolution of the sensor, we can choose the value of the dilation parameter α accordingly.

Using the representation (16) we can apply different imaging technics. For image enhancement we can use the unsharp masking (see [5], [6]), i.e. to subtract a blurred version of an image from the image itself. For the representation of original image $f(x,y)$ we can choose in (16) the interpolating kernels (dilation by $\alpha = 2$), but to get blurred version $f_b(x,y)$, we choose in (16) the dilated kernels with small parameter α, like $s_{H,1/2}$ in Figure 2. We can control the amount of unsharp masking choosing the parameter $a < 0$:

$$f_{usm}(x,y) = (1-a)f(x,y) + af_b(x,y).$$

Another well-known image enhancement method uses the derivatives of image. First derivatives in image processing are implemented using the magnitude of the gradient. The representation (16) gives us a natural way to implement derivatives. Indeed

$$f_x(x,y) := \sum_{j,k} f(j,k)s_1'(x-j)s_2(y-k),$$
$$f_y(x,y) := \sum_{j,k} f(j,k)s_1(x-j)s_2'(y-k).$$

Surprisingly, if we choose Hann kernel $s_1 = s_2 = s_H$ and $x, y \in \mathbb{Z}$, then the discrete convolution

$$f_x(p,q) \approx \sum_{j=p-1}^{p+1}\sum_{k=q-1}^{q+1} f(j,k)s_H'(p-j)s_H(q-k) \quad (17)$$

gives us the well-known Sobel filter (see [5], [6])

$$\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{pmatrix}.$$

Indeed, $s_H(k) = 0$ ($k \in \mathbb{Z}$) if $|k| > 1$ (see Figure 1) and we get $\frac{1}{4}(1,2,1)$. For s_H' we use the first 3 values only, i.e. $\frac{3}{8}(1,0,-1)$.

We can easily compute a directional derivative

$$f_\varphi(x,y) := \sum_{j,k} f(j,k)s_1'\Big((x-j)\cos\varphi - (y-k)\sin\varphi\Big)\times$$
$$\times s_2\Big((y-k)\cos\varphi + (x-j)\sin\varphi\Big),$$

To get the edges with different spatial frequency, we choose the dilation parameter (see Figure 1).

Second derivatives in image processing are implemented using the Laplacian. Using the representation (16) we get

$$\triangle f(x,y) := f_{xx}(x,y) + f_{yy}(x,y) =$$
$$\sum_{j,k} f(j,k)\big(s''(x-j)s(y-k) + s(x-j)s''(y-k)\big).$$

In image processing we use the derivatives for edge detection. Changing the dilation parameter α for the kernel $s_\alpha(t) = \alpha s(\alpha t)$ we can detect edges with different spatial frequencies.

In calculations we must use the truncated sampling series ($p, q \in \mathbb{Z}$)

$$(Sf)_{mn}(p,q) := \sum_{j=p-m}^{p+m}\sum_{k=q-n}^{q+n} f(j,k)s_1(p-j)s_2(q-k)$$

and have the truncation error. We can use kernels with finite support like the combinations of B-splines, considered in [4], to get rid of the truncation error, but in some

Figure 2: Unsharp mask with Hann kernel $s_{H,1/2} = \frac{1}{2}s_H(\frac{1}{2}t)$, $a = -1.7$.

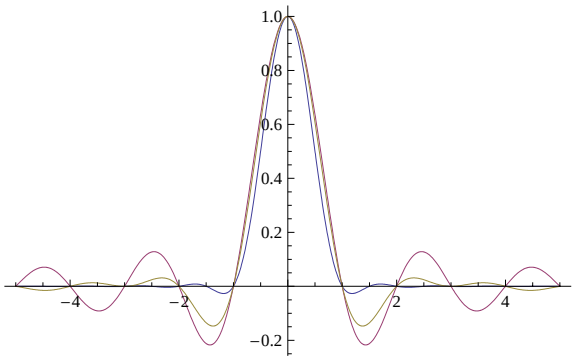


Figure 3: Hann kernel $\tilde{s}_H(t) = O(|t|^{-3})$, Lanczos kernel $s_{L,3}(t) = O(|t|^{-2})$ and $\mathrm{sinc}(t) = O(|t|^{-1})$.

cases other types of kernels are more suitable. For minimizing the truncation error the kernel $s(t)$ must decrease rapidly when $|t| \to \infty$. The sinc function does not belong even to L^1. Therefore using the kernels in form $s(t) = \theta(t)\,\mathrm{sinc}\,t$, where $\theta(t)$ is some window function (see [11]), is well-known. In most cases of we lose the important property (2) and do not get a generalized sampling series anymore. The kernels in our approach, i.e. kernels defined via Fourier transform of window functions, allow us to get good approximation properties and are rapidly decreasing. In Figure 3 we take the Hann kernel $\tilde{s}_H(t) = O(|t|^{-3})$ and compare it with the Lanczos kernel $s_{L,3}(t) = O(|t|^{-2})$, which is one of the most used kernels in imaging (see [16]). In the case of Blackman-Harris kernels (5), considered more precisely in [9], we have $s_{C,\mathbf{b}} = (|t|^{-2\ell-1})$ if for every $k = 0, \ldots, \ell - 1$

$$\sum_{j=0}^{m} j^{2k} b_j = 0.$$

We defined many rapidly decreasing kernels also in [8], [7], [10].

4. Acknowledgments

This research was supported by European Social Fund Funds Doctoral Studies and Internationalisation Programme DoRa, by the Estonian Science Foundation, grants 6943, 7033, and by the Estonian Min. of Educ. and Research, projects SF0132723s06, SF0140011s09. The second author wants to thank Archimedes Foundation for support.

References:

[1] H. H. Albrecht. A family of cosine-sum windows for high resolution measurements. In *IEEE International Conference on Acoustics, Speech and Signal Processing, Salt Lake City, Mai 2001*, pages 3081–3084. Salt Lake City, 2001.

[2] R. B. Blackman and J. W. Tukey. *The measurement of power spectra*. Wiley-VCH, New York, 1958.

[3] P. L. Butzer, G. Schmeisser, and R. L. Stens. An introduction to sampling analysis. In F Marvasti, editor, *Nonuniform Sampling, Theory and Practice*, pages 17–121. Kluwer, New York, 2001.

[4] P. L. Butzer, W. Splettster, and R. L. Stens. The sampling theorems and linear prediction in signal analysis. *Jahresber. Deutsch. Math-Verein*, 90:1–70, 1988.

[5] R. C. Gonzalez and R. E. Woods. *Digital Image Processing. Second Edition*. Prentice-Hall, 2002.

[6] B. Jähne. *Digital Image Processing: Concepts, Algorithms, and Scientific Applications*. Springer Verlag, Basel, Stuttgart, 1997.

[7] A. Kivinukk and G. Tamberg. On sampling operators defined by the Hann window and some of their extensions. *Sampling Theory in Signal and Image Processing*, 2:235–258, 2003.

[8] A. Kivinukk and G. Tamberg. Blackman-type windows for sampling series. *J. of Comp. Analysis and Applications*, 7:361–372, 2005.

[9] A. Kivinukk and G. Tamberg. On Blackman-Harris windows for Shannon sampling series. *Sampling Theory in Signal and Image Processing*, 6:87–108, 2007.

[10] A. Kivinukk and G. Tamberg. Interpolating generalized Shannon sampling operators, their norms and approximation properties. *Sampling Theory in Signal and Image Processing*, 8:77–95, 2009.

[11] R. J. Marks. *Fourier Analysis and Its Applications*. Oxford University Press, New York, 2009.

[12] E. Meijering and et al. Quantitative comparison of sinc-approximating kernels for medical image interpolation. In C. Taylor and A. Colchester, editors, *Medical Image Computing and Computer-Assisted Intervention*, pages 210–217. Springer, Berlin, 1999.

[13] S. M. Nikolskii. *Approximation of Functions of Several Variables and Imbedding Theorems*. Springer, Berlin, 1975. (Orig. Russian ed. Moscow, 1969).

[14] R. L. Stens. Sampling with generalized kernels. In J. R. Higgins and R. L. Stens, editors, *Sampling Theory in Fourier and Signal Analysis: Advanced Topics*. Clarendon Press, Oxford, 1999.

[15] M. Theis. Über eine interpolationsformel von de la Vallee-Poussin. *Math. Z.*, 3:93–113, 1919.

[16] K. Turkowski. Filters for common resampling tasks. In A. S. Glassner, editor, *Graphics Gems I*, pages 147–165. Academic Press, 1990.

Reconstruction of signals in a shift-invariant space from nonuniform samples

Junxi Zhao
College of Mathematics and Physics,
Nanjing university of posts and telecommunications,
Nanjing,21003, P.R. China
junxi_zhao@163.com

Abstract- In this paper, we consider the method for reconstructing a signal from a finite number of samples. In shift-invariant space framework, we derive an aproximately min-max optimal interpolator to to reconstruct a signal on an interval. An effective non-iterative algorithm for signal reconstruction is given also. Numerical examples show the effectiveness of the proposed method.

Index Terms–sampling, signal reconstruction, scaling function, shift-invariant space

Ⅰ. INTRODUCTION

The problem of signal reconstruction is pervasive in many areas of signal processing, such as in designs of nonuniform antenna arrays, sparse array beamforming, the restoration of signals with missing samples, image acquisition, etc [1-3]. The classical Shannon's sampling theorem was extended theoretically to general shift-invariant subspaces, and various generalized sampling theorems concerning band-limited and nonband-limited signals have been proposed [4-9]. However, the problem of reconstructing a continuous-time signal from its finite number of nonuniform samples is often encountered in practical applications, and truncating infinite reconstruction leads to errors.

Bandlimited and non bandlimited signals are often modeled by shift-invariant spaces. Some authors have proposed interpolation methods and iterative methods for reconstructing signals in shift-invariant spaces in the signal processing literatures[10-14]. A non-iterative reconstruction method is effective to reconstruct continuous-time signals from a finite number of samples by using a suitable interpolator. The Yen interpolator is well known to reconstruct band-limited signals in both minimal energy and least squares senses [13]. Some interpolation methods in shift-invariant spaces were given[9-10,12-14].

In this paper, we are interested in optimally constructing signals in a shift-invariant space from a finite number of nonuniform samples, and develop a new method for reconstructing continuous-time signal on a interval. The upper bound of reconstruction error is derived. We also propose a practical reconstruction algorithm. The method of signal reconstruction can be regarded as a generalization of Yen's in general shift-invariant spaces.

The paper is organized as follows: in Section Ⅱ we formulate the optimal reconstruction problem in shift-invariant spaces, and Section Ⅲ derives a new method to reconstruct a signal from a finite number of arbitrarily distributed samples. Section Ⅳ propose a practical algorithm for implementing the optimal signal reconstruction. In Section Ⅴ, some numerical examples of reconstructing signals demonstrate the effectiveness of the proposed method.

Ⅱ. THE PROBLEM FORMULATION

Throughout the paper, we focus on one-dimensional signals and denote the space of signals of finite energy on $\mathbb{R}$ by $\mathrm{L}^2(\mathbb{R})$. Let $\| f \|^2 = \int_{\mathbb{R}} | f(t) |^2 \, dt$ be the energy of a signal $f(t) \in L^2(\mathbb{R})$. Given K scaling functions $\varphi_1(t), \varphi_2(t), \cdots, \varphi_K(t) \in L^2(\mathbb{R})$, the shift-invariant space $V(\varphi_1, \varphi_2, \cdots, \varphi_K)$ is a Hilbert space defined as

$$V(\varphi_1, \varphi_2, \cdots, \varphi_K) = \mathrm{close}\Big\{ f(t) \in \mathrm{L}^2(\mathbb{R}) : f(t) = \sum_{k=1}^{K} \sum_{n} c_k(n) \varphi_k(t-n), \big(c_k(n)\big) \in l^2(\mathbb{Z}), k = 1, 2, \cdots, K \Big\}$$

We assume that $\{ \varphi_k(t-n) \,|\, 1 \le k \le K, n \in Z \}$ forms a frame of $V(\varphi_1, \varphi_2, \cdots, \varphi_K)$, i.e., there exist two constants $A > 0$ and $B < +\infty$ such that

$$A \| f \|^2 \le \sum_{k=1}^{K} \sum_{n} | (f, \varphi_k(t-n)) |^2 \le B \| f \|^2 \quad (1)$$

for any $f \in V(\varphi_1, \varphi_2, \cdots, \varphi_K)$.

To make the sampling of functions in $V(\varphi_1, \varphi_2, \cdots, \varphi_K)$ well-defined, we additionally assume that there exists a constant C such that

$$\sum_{k=1}^{K} \sum_{n} | \varphi_k(t-n) | < C \quad (2)$$

for any $t \in [0, 1]$. To see this, let $f(t) = \sum_{k=1}^{K} \sum_{n} c_k(n) \varphi_k(t-n)$ with $\big(c_k(n)\big) \in l^2(\mathbb{Z})$. From (2) and (3) it follows that

$$| f(t) | \le \left(\sum_{k=1}^{K} \sum_{n} | \varphi_k(t-n) |^2 \right)^{1/2} \left(\sum_{k=1}^{K} \sum_{n} | c_k(n) |^2 \right)^{1/2} \le C' \| f \|, \quad t \in R \quad (3)$$

where C' is a constant.

It is known from [8] that the assumption (3) implies that $V(\varphi_1, \varphi_2, \cdots, \varphi_K)$ is a reproducing kernel Hilbert space. For a function $f(t)$ in $V(\varphi_1, \varphi_2, \cdots, \varphi_K)$, we adopt the

following interpolator

$$\tilde{f}(t)=\sum_{m=1}^{M}h_m(t)f(t_m)$$

to reconstruct $f(t)$ on the interval containing sampling instants $t_1,\cdots,t_M$. In order to obtain an optimal interpolator, we discuss the optimization

$$\inf_{\mathbf{h}}\sup_{f}\frac{|f(t)-\tilde{f}(t)|^2}{\|f\|^2},$$

which yields a min-max type interpolator.

III. DERIVATION OF SIGNAL RECONSTRUCTION

Given M samples of a function $f(t)\in V(\varphi_1,\varphi_2,\cdots,\varphi_K)$ at nonuniform instants $t_1,t_2,\cdots,t_M\in[a,b]$, let $\tau\in[a,b]$ and $\tilde{f}(\tau)=\sum_{m=1}^{M}h_m(\tau)f(t_m)$. The optimal estimation $\tilde{f}(\tau)$ of $f(\tau)$ is determined by appropriate coefficients $h_m(\tau)$'s. So ,we study the following optimization

$$\inf_{\mathbf{h}(\tau)}\sup_{f\in V(\varphi_1,\cdots,\varphi_K)}\frac{|f(\tau)-\tilde{f}(\tau)|^2}{\|f\|^2} \tag{4}$$

Letting

$$f(t)=\sum_{n}\sum_{k=1}^{K}c_k(n)\varphi_k(t-n), \tag{5}$$

where, $c_k(n)=\left(f,\tilde{\varphi}_k(t-n)\right),\ k=1,2,\cdots,K,\ n\in Z$, and $\{\tilde{\varphi}_k(t-n)\,|\,1\le k\le K,n\in Z\}$ is the dual frame of $\{\varphi_k(t-n)\,|\,1\le k\le K,n\in Z\}$. We have

$$\begin{aligned}&|f(\tau)-\tilde{f}(\tau)|^2\\&=|\sum_{n}\sum_{k=1}^{K}c_k(n)[\varphi_k(\tau-n)-\sum_{m=1}^{M}h_m(\tau)\varphi_k(t_m-n)]|^2\\&\le\sum_{n}\sum_{k=1}^{K}|c_{k,n}(n)|^2\sum_{n}\sum_{k=1}^{K}|\varphi_k(\tau-n)-\sum_{m=1}^{M}h_m(\tau)\varphi_k(t_m-n)|^2\\&\le A^{-1}\|f\|^2\sum_{n}\sum_{k=1}^{K}|\varphi_k(\tau-n)-\sum_{m=1}^{M}h_m(\tau)\varphi_k(t_m-n)|^2.\end{aligned} \tag{6}$$

The above can furthermore be expressed explicitly in vector form as

$$|f(\tau)-\tilde{f}(\tau)|^2\le A^{-1}\|f\|^2\sum_{k=1}^{K}\|\mathbf{e}_{k,\tau}-\sum_{m=1}^{M}h_m(\tau)\mathbf{e}_{k,m}\|^2, \tag{7}$$

where $\mathbf{e}_{k,\tau}^T=(\varphi_k(\tau-n))_n$, $\mathbf{e}_{k,m}^T=(\varphi_k(t_m-n))_n$, $k=1,2,\cdots,K$. So, it follows that

$$\sup_{f}\frac{|f(\tau)-\tilde{f}(\tau)|^2}{\|f\|^2}\le A^{-1}\sum_{k=1}^{K}\|\mathbf{e}_{k,\tau}-\sum_{m=1}^{M}h_m(\tau)\mathbf{e}_{k,m}\|^2.$$

It can be seen that minimizing $E(\tau)=\sum_{k=1}^{K}\|\mathbf{e}_{k,\tau}-\sum_{m=1}^{M}h_m(\tau)\mathbf{e}_{k,m}\|^2$ can make $\tilde{f}(\tau)$ approximate $f(t)$ at τ .

To minimize $E(\tau)$, we further express $E(\tau)$ as

$$\begin{aligned}&E(\tau)=\sum_{k=1}^{K}\|\mathbf{e}_{k,\tau}-\sum_{m=1}^{M}h_m(\tau)\mathbf{e}_{k,m}\|^2=\\&\sum_{k=1}^{K}\|\mathbf{e}_{k,\tau}\|^2-2\sum_{k=1}^{K}\mathbf{e}_{k,\tau}^T\mathbf{A}_k\mathbf{H}(\tau)+\mathbf{H}^T(\tau)\left(\sum_{k=1}^{K}\mathbf{A}_k^T\mathbf{A}_k\right)\mathbf{H}(\tau).\end{aligned}$$

where $\mathbf{H}(\tau)=(h_1(\tau),h_2(\tau),\cdots,h_M(\tau))^T$ and $\mathbf{A}_k=\left(\mathbf{e}_{k,1},\mathbf{e}_{k,2},\cdots,\mathbf{e}_{k,M}\right)$ for $k=1,2,\cdots,K$. It is followed that

$$\frac{\partial E(\tau)}{\partial\mathbf{H}(\tau)}=-2\sum_{k=1}^{K}\mathbf{A}_k^T\mathbf{e}_{k,\tau}+2\sum_{k=1}^{K}\mathbf{A}_k^T\mathbf{A}_k\mathbf{H}(\tau). \tag{8}$$

Therefore, solving $\frac{\partial E(\tau)}{\partial\mathbf{H}(\tau)}=0$ yields that

$$\mathbf{H}(\tau)=\left(\sum_{k=1}^{K}\mathbf{A}_k^T\mathbf{A}_k\right)^{-1}\sum_{k=1}^{K}\mathbf{A}_k^T\mathbf{e}_{k,\tau}. \tag{9}$$

and the minimal error between $f(\tau)$ and $\tilde{f}(\tau)$ is given by

$$\begin{aligned}&|f(\tau)-\tilde{f}(\tau)|\le r(\tau)=\\&\sqrt{A^{-1}}\,\|f\|\left[\sum_{k=1}^{K}\|\mathbf{e}_{k,\tau}-\mathbf{A}_k(\sum_{l=1}^{K}\mathbf{A}_l^T\mathbf{A}_l)^{-1}\sum_{m=1}^{K}\mathbf{A}_m^T\mathbf{e}_{m,\tau}\|^2\right]^{1/2}.\end{aligned} \tag{10}$$

Note that the matrix $\sum_{k=1}^{K}\mathbf{A}_k^T\mathbf{A}_k$ is invertible when the samples $\{f(t_m)\}_{m=1}^{M}$ is indepentant, that is, there doesn't exist one sample in $\{f(t_m)\}_{m=1}^{M}$ which can be expressed by the others for each $f\in V(\varphi_1,\cdots,\varphi_M)$.

Letting $\mathbf{A}_k=(a_{ij}^{(k)})$ with $a_{ij}^{(k)}=\varphi_k(t_j-i)$ and $\mathbf{X}=(x_{ij})=\left(\sum_{k=1}^{K}\mathbf{A}_k^T\mathbf{A}_k\right)^{-1}$, we then rewrite the optimal interpolating vector as

$$\begin{aligned}&\mathbf{H}(\tau)=\left(h_1(\tau),h_2(\tau),\cdots,h_M(\tau)\right)^{\mathrm{T}}=\mathbf{X}\sum_{k=1}^{K}\mathbf{A}_k^T\mathbf{e}_{k,\tau}=\\&\sum_{k=1}^{K}\mathbf{X}(\sum_{n}\varphi_k(t_1-n)\varphi_k(\tau-n),\cdots,\sum_{n}\varphi_k(t_M-n)\varphi_k(\tau-n))^T\\&=(\sum_{k=1}^{K}\sum_{l=1}^{M}\sum_{n}x_{1l}\varphi_k\left(t_l-n\right)\varphi_k(\tau-n),\cdots,\\&\qquad\sum_{k=1}^{K}\sum_{l=1}^{M}\sum_{n}x_{Ml}\varphi_k\left(t_l-n\right)\varphi_k(\tau-n))^T.\end{aligned} \tag{11}$$

So, the optimal reconstruction of $f(t)$ can be expressed as

$$\begin{aligned}\tilde{f}(t)&=\sum_{m=1}^{M}f(t_m)h_m(t)\\&=\sum_{m=1}^{M}f(t_m)\left[\sum_{k=1}^{K}\sum_{l=1}^{M}\sum_{n}x_{ml}\varphi_k\left(t_l-n\right)\varphi_k(t-n)\right].\end{aligned} \tag{12}$$

Let us take a look at the case $K=1$ in detail. Given M samples of $x(t)\in V(\varphi)$ at instants $t_1,t_2,\cdots,t_M$ for a proper scaling function $\varphi(t)\in\mathrm{L}^2(\mathbb{R})$, we can express the optimal interpolating vector as $\mathbf{H}(\tau)=(A^TA)^{-1}A^T\mathbf{e}_\tau$,

where $A=(\mathbf{e}_1,\mathbf{e}_2,\cdots,\mathbf{e}_M)$ and $\mathbf{e}_\tau=(\varphi(\tau-n))_n^T$, $\mathbf{e}_m=(\varphi(t_m-n))_n^T$ for $m=1,2,\cdots,M$. It is easy to see that the optimal interpolating vector $\mathbf{H}(\tau)$ is exactly the orthogonal projection of vector $\mathbf{e}_\tau$ onto the subspace spanned by $\mathbf{e}_1,\mathbf{e}_2,\cdots,\mathbf{e}_M$ and hence from (12) $\tilde{x}(t)\in V(\varphi)$ with $x(t_m)=\tilde{x}(t_m)$ for $m=1,2,\cdots,M$. This implies that the reconstructed signal best fits the sampling data. Especially, for $\sigma>0$ and $\varphi(t)=\mathrm{sinc}(\sigma t)$, the optimal reconstruction $\tilde{x}(t)$ of $x(t)\in V(\varphi)$ from samples $x(t_1),x(t_2),\cdots,x(t_M)$ is also band-limited to σ with $x(t_m)=\tilde{x}(t_m)$ for $m=1,2,\cdots,M$. It is easy to show that the interpolator obtained here is just Yen's for band-limited signals.

Ⅳ. ALGORITHM AND DISCUSSION

In the previous section we have derived an interpolator for signal reconstruction. However, because computing the min-max interpolator requires calculating the inverse of a matrix with possibly larger dimension, the reconstruction formula (12) wound be unfeasible when the number of samples is much large. To circumvent this problem, we reshape (12) as

$$\tilde{f}(\tau)=\sum_n\left(\sum_{l=1}^{M}\sum_{k=1}^{K}\sum_{m=1}^{M}f(t_m)x_{ml}\varphi_k(t_l-n)\right)\varphi_k(t-n). \quad (13)$$

From this, a non-iterative reconstruction algorithm can be given as follows.

Algorithm:

(1) Let $\mathbf{f}=(f(t_1),f(t_2),\cdots,f(t_M))^{\mathrm{T}}$,
$\mathbf{E}_k(t)=(\cdots,\varphi_k(t-n),\varphi_k(t-n+1),\cdots)^{\mathrm{T}}$,
$\mathbf{A}_k=(b_{mn}^{(k)})$ with $b_{mn}^{(k)}=\varphi_k(t_m-n)$ for $i=1,2,\cdots,M$ and $k=1,2,\cdots,K$;

(2) Compute $\mathbf{T}=\sum_{k=1}^{K}\mathbf{A}_k\mathbf{A}_k^{\mathrm{T}}$

(3) Solve $\mathbf{Th}=\mathbf{f}$;

(4) $\tilde{f}(t)=\mathbf{h}^{\mathrm{T}}\sum_{k=1}^{K}\mathbf{A}_k\mathbf{E}_k(t)$.

The most crucial step is solving the equation system of equations $\mathbf{Th}=\mathbf{f}$. This can be done effectively by computing the Cholesky factorization of the matrix $\mathbf{T}$ when $\mathbf{T}$ is invertible. In fact, the cholesky factorization of $\mathbf{T}$ gives a upper triangular matrix $\mathbf{S}$ such that $\mathbf{T}=\mathbf{S}^T\mathbf{S}$. Then the solution of the system $\mathbf{Th}=\mathbf{f}$ can be obtained by sequentially solving the systems $\mathbf{S}^T\mathbf{b}=\mathbf{f}$ and $\mathbf{Sh}=\mathbf{b}$ by Gaussian elimination. This procedure is faster and more robust than directly computing the inverse of $\mathbf{T}$. When $\mathbf{T}$ is not invertible, the equation can be solved effectively by the least squares method.

Note that although the proposed method has no restriction on sampling locations, the obtained reconstruction error is strongly related to the sampling pattern. As pointed in [18], the reconstruction errors are smaller in the neighborhood of the sampling instants. So, the quality of reconstruction should be evaluated in a pointwise manner. From (10) we know that the min-max reconstruction error is pointwise upper-bounded by

$$r(\tau)\le A^{-1/2}\|f\|\left[\sum_{k=1}^{K}\|\mathbf{e}_{k,\tau}-\mathbf{A}_k(\sum_{l=1}^{K}\mathbf{A}_l^T\mathbf{A}_l)^{-1}\sum_{m=1}^{K}\mathbf{A}_m^T\mathbf{e}_{m,\tau}\|^2\right]^{1/2} \quad (14)$$

and it can estimate the quality of reconstruction when the sampling instants are known.

Ⅴ. DEMONSTRATIVE EXAMPLES

Some numerical examples are given to demonstrate the performance of the proposed method, where signals are selected randomly in shift-invariant subspaces, and the sampling instants are generated by adding random perturbation, distributed uniformly in the interval $[-u,u]$ for $u>0$, to each equally-spaced sampling instant, i.e., the sampling instants are $mT+u_m$, where u_m randomly distributes in $[-u,u]$ for $m=1,2,\cdots,M$.

Example 1 For the first example, we choose arbitrarily a signal of band $[-\pi,\pi]$. We reconstruct it on $[0,40]$ from 42 samples. The average sampling period is $T=0.995$ s and $u=0.7T$. It is clear that the average sampling period is almost critical. The reconstructed signal, its reconstruction from its nonuniform samples and the errors between the original signal and its reconstructions were plotted in Fig.1. From this experiment, it can be seen that under such a relaxed condition, the reconstruction of a signal is quite satisfying,

Example 2 For non band-limited signals, we choose the cubic spline [19] as a scaling function, and randomly choose a signal in the shift-invariant space. The average sampling period is $T=1.0$s and the maximum of irregular perturbation is $u=0.5T$. The signal to be reconstructed, its reconstruction and the reconstruction error (in dB) were plotted in Fig.2, respectively. Note that, although the sampling density is much lower than that estimated in [5](see the examples therein for details), the quality of signal reconstruction is still considerably high. In addition, although the cubic spline is supported compactly, the method given in [10] could not be used in this case because the maximal gap between adjacent sampling instants is too large. In fact, the local reconstruction methods in [10] required the condition that the maximal gap of the sampling instants must be less 1 and the number of the samples must be bigger than the length of the reconstruction interval, but our method doesn't rely on any sampling condition. In contrast to the results in [10], we also give another example to show the performance of the proposed method in Fig.3. In this example, we chose the same scaling function, a Gaussian function, and the similar sampling condition as in [10].

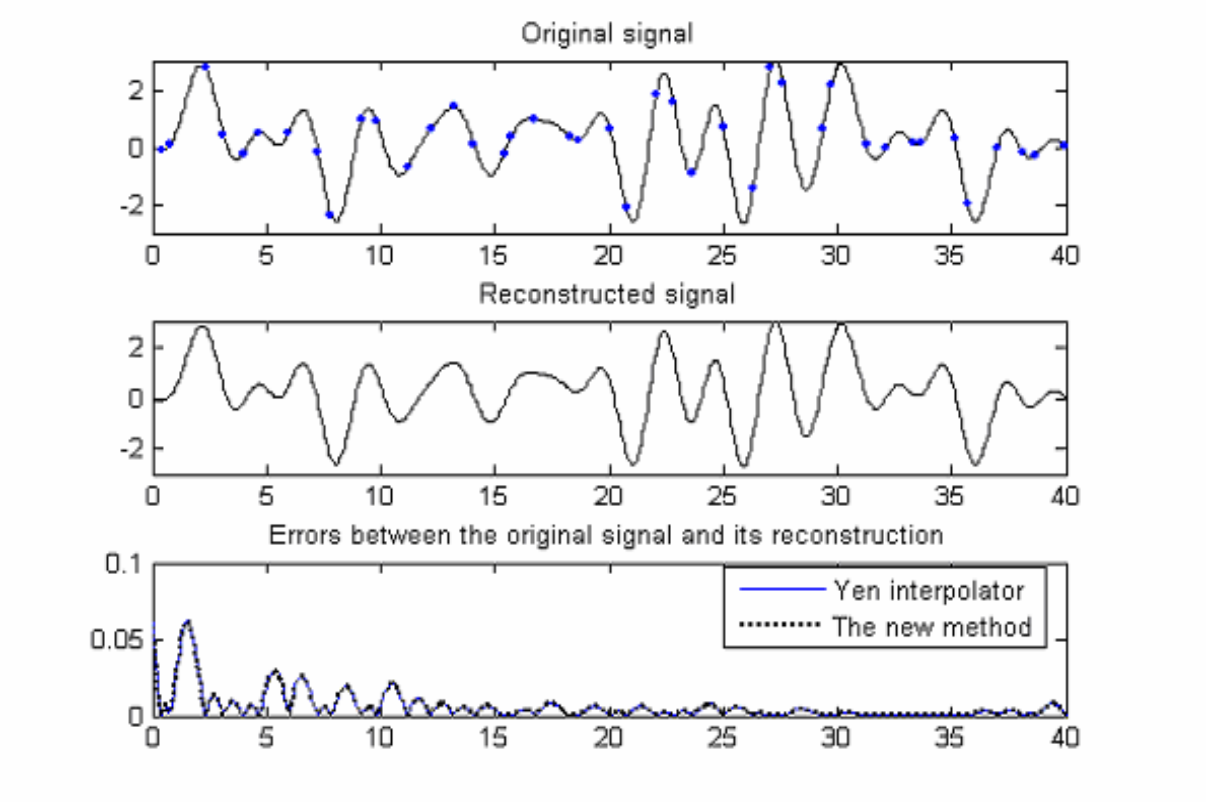


Fig. 1 Top : original signal and sampling points marked by dots; middle: reconstructed signal obtained by the proposed method; bottom: normalized errors between the original signal the its reconstruction obtained by the proposed method and Yen interpolator.

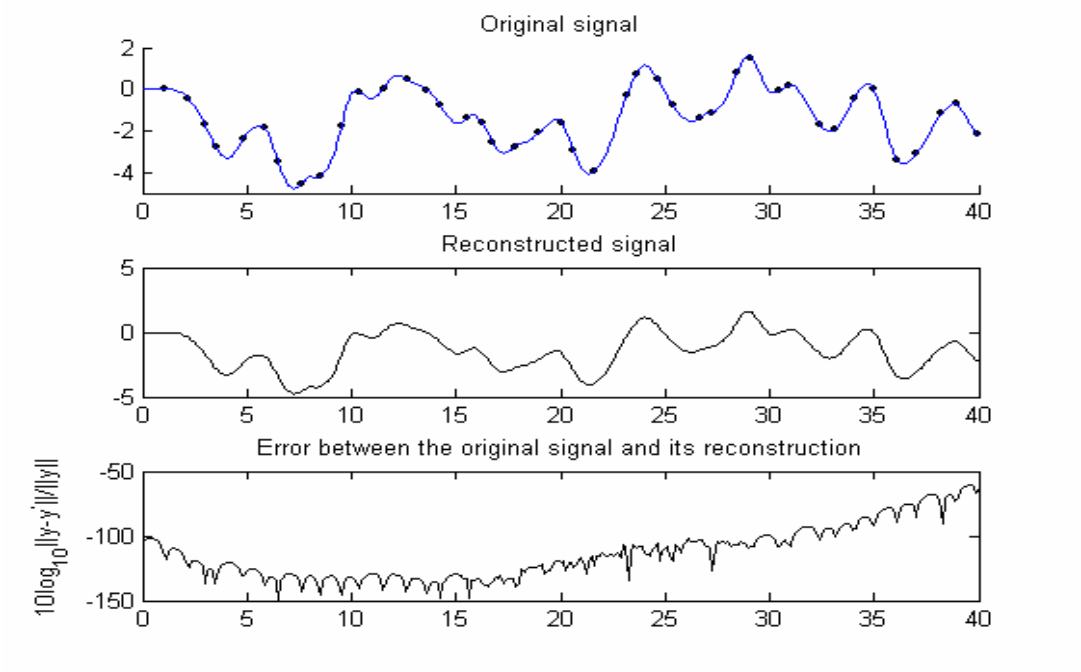


Fig. 2 original signal with sampling points marked by stars, reconstructed signal obtained by the proposed method, normalized error(in dB) between the original signal the its reconstruction.

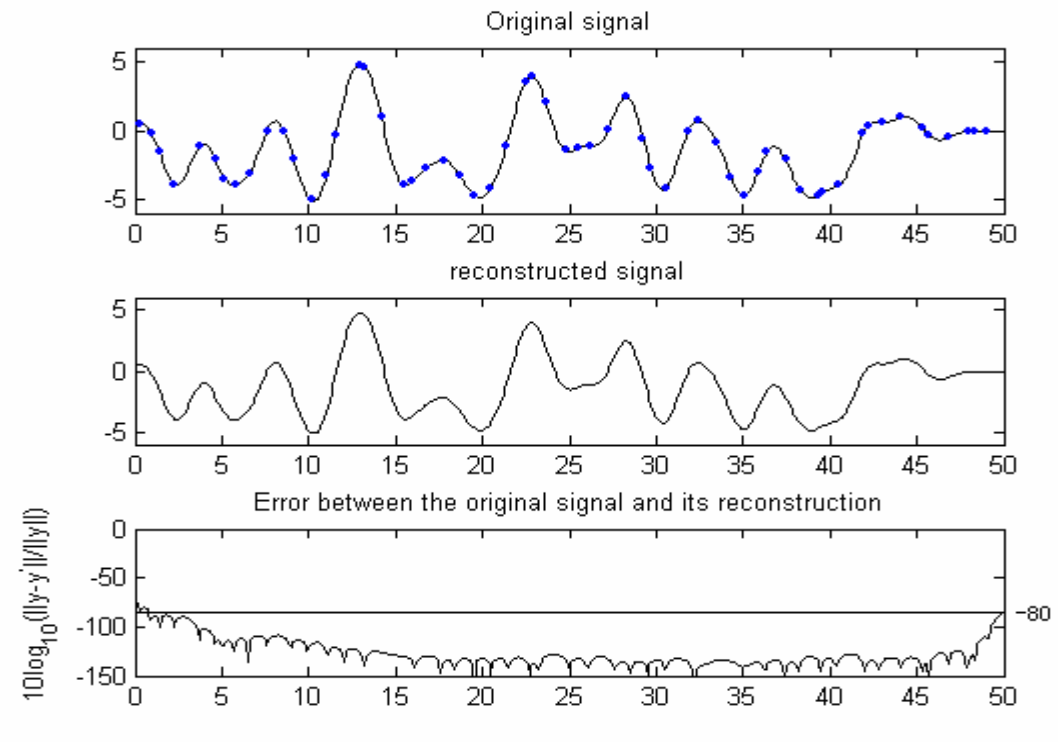


Fig. 3 original signal with sampling points marked by stars, reconstructed signal obtained by the proposed method, normalized error(in dB) between the original signal the its reconstruction with scaling function $\varphi(t) = e^{-t^2/2\sigma^2}$, $\sigma = 0.81$, and sampling density 0.85.

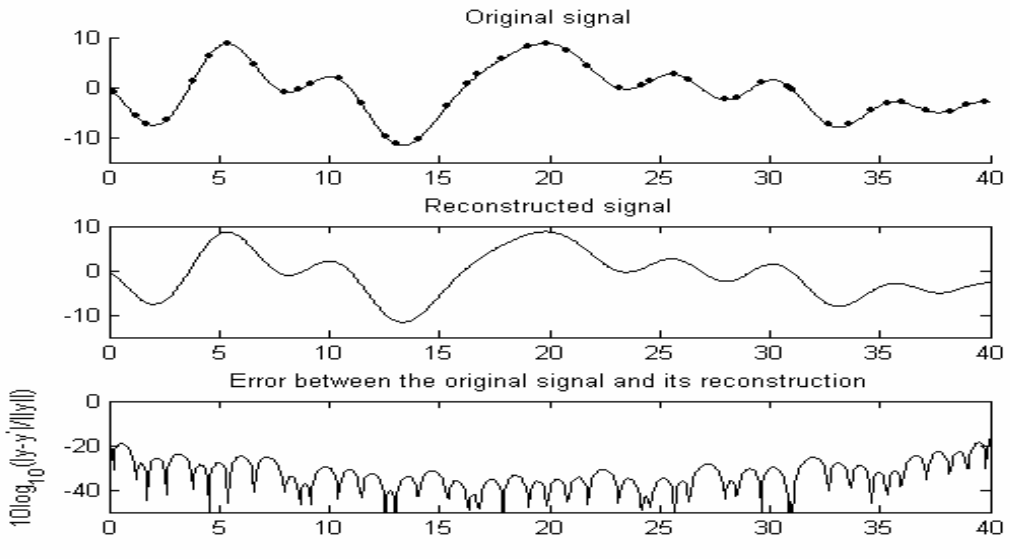


Fig. 4 original signal with sampling points marked by stars, reconstructed signal obtained by the proposed method, normalized error(in dB) between the original signal the its reconstruction.

Example 3. Finally, we select two functions $\varphi_1(t) = a_1 e^{-t^2/4}$ and $\varphi_2(t) = a_2(t+t^3)e^{-t^2/4}$ as scaling functions, where a_1 and a_2 are normalized constants. Here the average sampling period is $T = 0.8$ s and the maximum of irregular perturbation $u = 0.6T$. The simulation results were showed in Fig.4. This example also indicates the feasibility of the proposed method for signal reconstruction in a shift-invariant spaced with several scaling functions

Ⅵ. CONCLUSION

The proposed method of reconstructing a signal from its finite nonuniform samples has the following advantages: (a) the method doesn't require the usual hypotheses on the maximal gap between adjacent sampling instants and the compactness of the scaling functions of the shift-invariant space as in the literature, and therefore can be applied in various shift-invariant spaces with sampling locations distributed arbitrarily. (b) The reconstruction error function as sensitivity functions [17] can measure the quality of the reconstruction prior to the practical implementation. (c) the method can be used effectively in a multi-wavelet space and can be extended straightforward to two-dimensional spaces. However, our method does not incorporate the case when samples are noisy, which we will investigate in future.

REFERENCES

[1] S. D. Berger, "Nonuniform sampling reconstruction Applied to sparse array beamforming," *Proc. IEEE Radar Conf.* 2002, pp.98–103, 2002,

[2] D. S. Early and D.G. Long, "Image reconstruction and enhanced resolution imaging from irregular samples," *IEEE Trans. on Geoscience and Remote Sensing*, vol.39, pp.291–302, 2001.

[3] R. Stasinski and J. Konrad, "POCS-based image reconstrunction from nonuniform samples," http://iss.bu.edu/jkonrad/Publications/local/cpapers/Stas00icip.pdf.

[4] G. G. Water, "A sampling theorem for wavelet subspaces," *IEEE Trans. on Inform. Theory*, vol. 38, pp.881-884, 1992.

[5] Wen Chen, S. Itoh and J. Shiki, "On Sampling in Shift Invariant Spaces," *IEEE Trans. on Information Theory*, vol.48, pp.2802–28010, 2002.

[6] I. Djokovic and P.P. Vaidyanathan, "Generalized sampling theorem in multiresolution subspaces," *IEEE,* Trans. *on Signal Process*, vol.45, pp.583–599, 1997.

[7] I. W. Selesnick, "Interpolating multiwavelet bases and sampling theory," *IEEE Trans. on Signal Process*, vol.47, pp.1615–11621, 1999.

[8] A. Aldroubi and K. Grochenig, "Nonuniform sampling and reconstruction in shift-invariant spaces", *SIAM Rev.*, 2001, no.4, pp.585-620.

[9] C. Ford and D.M. Etter, "Wavelet basis reconstruction of nonuniformly sampled data", *IEEE Trans on Circuits* and *Systems II*, vol.45, pp.1165–1168, 1998.

[10] K. Grochenig and H. Schwab, "Fast local reconstruction methods for nonuniform sampling in shift-invariant spaces," *SIAM Journal of Matrix Analysis and Applications*, vol.24 , no.4, pp.899-913, 2003.

[11] A. Aldroubi and H. Feichtinger, "Complete iterative reconstruction algorithms for irregular sampled data in spline-like spaces", in *IEEE Acoustics, Speech and Signal Process International Conf.(ICASSP-97)*, vol.3, pp.1857-1860, 1997.

[12] P. J. S. G. Ferreira, "Noniterative and faster iterative methods for interpolation and extrapolation, " *IEEE Trans. on Signal Processing*, vol. 42(11), pp. 3278-3282, 1994.

[13] Hyeokho Choi and D. C. Munson, "Analysis and Design of minimax-optimal interpolators", *IEEE Trans. on Signal Processing*, vol.46, pp.1571–1579, 1998.

[14] Y. Rolain, J. Schoukens and G. Vandersreen, "Signal reconstruction for non-equidistant finite length sample sets: a KIS approach," *IEEE Trans. on Instrument and Measurement*, vol.47, no.5, pp.1046-1052,1998.

[15]I. A. Aldroubi and M. Unser, "Sampling procedures in function spaces and asymptotic equivalence with Shannon sampling," *Numer. Funct. Anal. Optimiz.*, vol.15, pp. 1-21,1994.

[16] P.P. Vaidyanathan, "Generalizations of the sampling theorem: Seven decades after Nyquist", *IEEE, Trans. on Circuits and Systems I: Fundamental Theory and Applications*, vol.48, pp.1094–1109, 2001.

[17] R. G. Shenoy and T.W. Parks, "An optimal recovery approach to interpolation", *IEEE, Trans. on Signal* Processing, Vol.40, pp.1987-1996, 1992.

[18]A. Tarczynski, "Sensitivity of signal reconstruction", *IEEE Signal Processing Letter*, vol.4, pp.192–194, 1997.

[19]C. K. Chui, An Introduction to wavelets, Academic Press, Inc. 1992.

Spline Interpolation in Piecewise Constant Tension

Masaru Kamada[(1)] and Rentsen Enkhbat[(2)]

(1) Ibaraki University, Hitachi, Ibaraki 316 8511, Japan.
(2) National University of Mongolia, P. O. Box 46/635, Ulaanbaatar, Mongolia.
kamada@mx.ibaraki.ac.jp, renkhbat46@ses.edu.mn

Abstract:

Locally supported splines in tension are constructed where the tension, which has ever been constant over the entire domain, is allowed to change at sampling points.

1. Introduction

A cubic spline gives the interpolation of data that minimizes the square integral of its second derivative [3, 5, 9] and is crowned as the smoothest interpolation in this sense. A linear spline gives the piecewise linear interpolation that is most straight but nonsmooth. The linear spline is characterized as minimizing the square integral of its first derivative [3, 9]. A *spline in tension* [1, 10] was devised as a generalization of those two splines. It minimizes the integral of a weighted sum of the squared second derivative and the squared first derivative. By increasing the weight called *tension*, we can make a spline in tension approach the most straight linear spline while retaining smoothness similar to that of the cubic spline.

The spline in tension has been known for more than 40 years. It has been extended even to the multidimensional cases [2, 7] and is now supported by a standard software library [8]. But the tension has ever been a single constant over the entire domain.

In this paper, we look at the splines as the output of a linear dynamical system with a series of delta functions input. That is the same way as how the exponential splines and their locally supported basis were successfully constructed in [12, 13]. In addition, attending to that the linear dynamical system theory [6] allows for time-varying dynamical parameters, we shall place different tension in each sampling interval. Then we will obtain locally supported splines in piecewise constant tension that can change the interpolation characteristics from a sampling interval to another.

2. Preliminaries

A spline f in tension interpolating the data $\{(t_k, f_k)\}_{k=-\infty}^{\infty}$ given at strictly increasing sampling points $(\cdots < t_{-2} < t_{-1} < t_0 < t_1 < t_2 < \cdots)$ on the real line is defined as the twice-differentiable function that minimizes the integral of a weighted sum

$$\int_{-\infty}^{\infty} (f^{(2)}(t))^2 + p(t)^2 (f^{(1)}(t))^2 \, dt \tag{1}$$

of its squared second derivative $f^{(2)}$ and squared first derivative $f^{(1)}$ subject to the constraints

$$f(t_k) = f_k, \quad k = 0, \pm 1, \pm 2, \cdots . \tag{2}$$

In the case $p = 0$, f is identical with the cubic spline [3, 5, 9]. By increasing p, f approaches the most straight linear spline as if the curve were pulled from both ends. That is why p is called *tension* [1, 10].

The tension p has originally been a single constant over the entire domain [10]. We shall now relax the tension to be a non-negative constant in each sampling interval, i.e.,

$$p(t) = p_k \geq 0, \ \text{ for } \ t \in [t_k, t_{k+1}), \tag{3}$$

which can change at the sampling points.

By the calculus of variation, the minimization problem is reduced to solution of the Euler-Lagrange differential equation

$$f^{(4)}(t) - 2p(t)p^{(1)}(t)f^{(1)}(t) - p(t)^2 f^{(2)}(t) = w(t), \tag{4}$$

where w is a series of the Dirac delta functions

$$w(t) = \sum_{n=-\infty}^{\infty} w_n \delta(t - t_n)$$

to be determined so that (2) holds good. We do not have, however, a practical means to decide the coefficients $\{w_n\}$ for given $\{(t_k, f_k)\}$.

In practice, it is convenient to have locally supported functions $\{y_n\}$ satisfying

$$y_n(t) = 0, \ \text{ for } \ t \notin [t_n, t_{n+4}] \tag{5}$$

of which linear combination

$$f(t) = \sum_{n=-\infty}^{\infty} c_n y_n(t) \tag{6}$$

represents any possible f. This y_n can be constructed by

$$y_n^{(4)}(t) - 2p(t)p^{(1)}(t)y_n^{(1)}(t) - p(t)^2 y_n^{(2)}(t) = u_n(t) \tag{7}$$

for some appropriately chosen

$$u_n(t) = \sum_{l=0}^{4} u_{l,n} \delta(t - t_{n+l}) \tag{8}$$

as long as the sampled data system (7) with the impulse input (8) is completely controllable [4]. Once we obtain $y_n(t)$, we have only to determine the coefficients $\{c_n\}$ by the linear equations

$$f_k = \sum_{n=-\infty}^{\infty} c_n y_n(t_k), \quad k = 0, \pm 1, \pm 2, \cdots$$

from $\{(t_k, f_k)\}$. Although infinitely many coefficients and data are involved in the equations, we can solve the linear equations for finitely many $\{c_n\}$ from finitely many $\{(t_k, f_k)\}$ in practice because $\{y_n\}$ are locally supported.

3. Construction of locally supported splines in piecewise constant tension

A state-space representation of the differential equation (7) is

$$\boldsymbol{x}_n^{(1)}(t) = F(t)x_n(t) + \boldsymbol{g}u_n(t), \;\; y_n(t) = \boldsymbol{h}\boldsymbol{x}_n(t), \quad (9)$$

where

$$F(t) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 2p(t)p^{(1)}(t) & p(t)^2 & 0 \end{bmatrix},$$

$$\boldsymbol{x}_n(t) = \begin{bmatrix} y_n \\ y_n^{(1)} \\ y_n^{(2)} \\ y_n^{(3)} \end{bmatrix}, \; \boldsymbol{g} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \; \boldsymbol{h} = [1\,0\,0\,0]. \quad (10)$$

The state $\boldsymbol{x}_n$ can be expressed by

$$\boldsymbol{x}_n(t) = \Phi(t, v)\boldsymbol{x}_n(v) + \int_v^t \Phi(t, \tau)\boldsymbol{g}u_n(\tau)\, d\tau, \quad (11)$$

for any real numbers t and v, in terms of the state-transition matrix function Φ and the input u_n [11].

Since $u_n(t) = 0$ for $t \notin \{t_n, t_{n+1}, t_{n+2}, t_{n+3}, t_{n+4}\}$, it follows from (11) that

$$\boldsymbol{x}_n(t) = \begin{cases} \boldsymbol{0}, & t < t_n \\ \Phi(t, t_{n+l}\text{+0})\boldsymbol{x}_n(t_{n+l}\text{+0}), & \\ \qquad t_{n+l} < t < t_{n+l+1}, (l = 0, 1, 2, 3) & \\ \Phi(t, t_{n+4}\text{+0})\boldsymbol{x}_n(t_{n+4}\text{+0}), & t_{n+4} < t. \end{cases} \quad (12)$$

Because of the top and bottom lines of (12), $y_n = \boldsymbol{h}\boldsymbol{x}_n$ is locally supported as (5) if $\boldsymbol{x}_n(t_{n+4}\text{+0}) = \boldsymbol{0}$. In order to avoid the trivial case $u_{0,n} = u_{1,n} = u_{2,n} = u_{3,n} = u_{4,n} = 0$ that would result in $u_n \equiv y_n \equiv 0$, let us fix one of them as $u_{0,n} = 1$. Then the problem of constructing a locally supported y_n becomes a dead-beat control problem of finding $u_{1,n}$, $u_{2,n}$, $u_{3,n}$, and $u_{4,n}$ that make the terminal state dead as

$$\boldsymbol{x}_n(t_{n+4}\text{+0}) = \boldsymbol{0}. \quad (13)$$

Once the terminal state is controlled to $\boldsymbol{0}$, it will stay at $\boldsymbol{0}$ forever for $t > t_{n+4}$ without any beats.

We shall consider two types of state transitions: (i) Those within each sampling interval (t_{n+l}, t_{n+l+1}), and (ii) one across each sampling point t_{n+l}.

(i) In the open interval (t_{n+l}, t_{n+l+1}), (11) with $v = t_{n+l}$+0 is reduced to

$$\boldsymbol{x}_n(t) = \Phi(t, t_{n+l}\text{+0})\boldsymbol{x}_n(t_{n+l}\text{+0}), \;\; l = 0, 1, 2, 3 \quad (14)$$

because $u_n(t) = 0$ for $t \in (t_{n+l}, t_{n+l+1})$. Besides, $F(t)$ in (10) is a constant matrix

$$F(t) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & p_{n+l}^2 & 0 \end{bmatrix} \quad (15)$$

because of (3) so that we can calculate the state-transition matrix by the matrix exponential function [11] as follows:

$$\begin{aligned} & \Phi(t, t_{n+l}\text{+0}) \\ = \;& e^{\int_{t_{n+l}+0}^{t} F(v)\, dv} \\ = \;& \begin{cases} \begin{bmatrix} 1 & t-t_{n+l} & \frac{(t-t_{n+l})^2}{2} & \frac{(t-t_{n+l})^3}{6} \\ 0 & 1 & t-t_{n+l} & \frac{(t-t_{n+l})^2}{2} \\ 0 & 0 & 1 & t-t_{n+l} \\ 0 & 0 & 0 & 1 \end{bmatrix} & \\ \qquad\qquad \text{if } p_{n+l} = 0 & \\ \begin{bmatrix} 1 & t-t_{n+l} & \frac{\cosh(p_{n+l}(t-t_{n+l}))-2}{p_{n+l}^2} & \\ 0 & 1 & \frac{\sinh(p_{n+l}(t-t_{n+l}))}{p_{n+l}} & \\ 0 & 0 & \cosh(p_{n+l}(t-t_{n+l})) & \\ 0 & 0 & p_{n+l}\sinh(p_{n+l}(t-t_{n+l})) & \end{bmatrix} & \\ \begin{bmatrix} \frac{\sinh(p_{n+l}(t-t_{n+l}))-2p_{n+l}(t-t_{n+l})}{p_{n+l}^3} \\ \frac{\cosh(p_{n+l}(t-t_{n+l}))-2}{p_{n+l}^2} \\ \frac{\sinh(p_{n+l}(t-t_{n+l}))}{p_{n+l}} \\ \cosh(p_{n+l}(t-t_{n+l})) \end{bmatrix} & \\ \qquad\qquad \text{if } p_{n+l} > 0. & \end{cases} \end{aligned} \quad (16)$$

In the special case that $t = t_{n+l+1}$-0, we have the state transition from $\boldsymbol{x}_n(t_{n+l}\text{+0})$ to $\boldsymbol{x}_n(t_{n+l+1}\text{-0})$ as follows:

$$\boldsymbol{x}_n(t_{n+l+1}\text{-0}) = \Phi(t_{n+l+1}\text{-0}, t_{n+l}\text{+0})\boldsymbol{x}_n(t_{n+l}\text{+0}), \quad l = 0, 1, 2, 3. \quad (17)$$

The matrix $\Phi(t_{n+l+1}\text{-0}, t_{n+l}\text{+0})$ can be evaluated by the right hand side of (16) with t replaced by t_{n+l+1}.

(ii) The state transition from $\boldsymbol{x}_n(t_{n+l}\text{-0})$ to $\boldsymbol{x}_n(t_{n+l}\text{+0})$ across the sampling point t_{n+l}, $(l = 0, 1, 2, 3, 4)$ finds a trouble that $F(t)$ in (10) contains a derivative of the function p being discontinuous at t_{n+l} as defined by (3). We had better consider this transition by way of the original differential equation (7). An equivalent form of (7) is

$$y_n^{(4)}(t) - \frac{d}{dt}\left(p(t)^2 y_n^{(1)}(t)\right) = u_n(t) \quad (18)$$

and its integration gives

$$y_n^{(3)}(t) = p(t)^2 y_n^{(1)}(t) + \int_{t_n-0}^{t} u_n(\tau)d\tau + c, \quad (19)$$

where c is an integral constant. Substituting t_{n+l}+0 and t_{n+l}-0 for t of (19), we have

$$\begin{aligned} y_n^{(3)}(t_{n+l}\text{+0}) = \;& p(t_{n+l}\text{+0})^2 y_n^{(1)}(t_{n+l}\text{+0}) \\ & + u_{0,n} + \cdots + u_{n+l,n} + c \end{aligned} \quad (20)$$

and

$$y_n^{(3)}(t_{n+l}\text{-}0) = p(t_{n+l}\text{-}0)^2 y_n^{(1)}(t_{n+l}\text{-}0) + u_{0,n} + \cdots + u_{n+l-1,n} + c, \quad (21)$$

respectively. Recall that the spline in tension is sought among the twice-differentiable functions and attend to the definition (3) of p. Then we can reduce (20) and (21) to

$$y_n^{(3)}(t_{n+l}\text{+}0) = p_{n+l}^2 y_n^{(1)}(t_{n+l}) + u_{0,n} + \cdots + u_{n+l,n} + c \quad (22)$$

and

$$y_n^{(3)}(t_{n+l}\text{-}0) = p_{n+l-1}^2 y_n^{(1)}(t_{n+l}) + u_{0,n} + \cdots + u_{n+l-1,n} + c, \quad (23)$$

respectively. Subtracting (23) from (22), we have

$$y^{(3)}(t_{n+l}\text{+}0) - y^{(3)}(t_{n+l}\text{-}0) = (p_{n+l}^2 - p_{n+l-1}^2) y^{(1)}(t_{n+l}) + u_{l,n}, \quad (24)$$

which tells how to update the state variable $y^{(3)}$ at t_{n+l} and implies that the other state variables $y^{(2)}$, $y^{(1)}$, and y are continuous at t_{n+l}. So we can write the state transition across the sampling point t_{n+l} as follows:

$$\boldsymbol{x}_n(t_{n+l}\text{+}0) = \Phi(t_{n+l}\text{+}0, t_{n+l}\text{-}0)\boldsymbol{x}_n(t_{n+l}\text{-}0) + \boldsymbol{g}u_{l,n}, \quad l = 0,1,2,3,4, \quad (25)$$

where

$$\Phi(t_{n+l}\text{+}0, t_{n+l}\text{-}0) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & p_{n+l}^2 - p_{n+l-1}^2 & 0 & 1 \end{bmatrix}. \quad (26)$$

The two types of state transitions (17) and (25) can be combined into the recurrence formulae

$$\begin{aligned} \boldsymbol{x}_n(t_n\text{+}0) &= \boldsymbol{g}u_{0,n} = \boldsymbol{g}, \\ \boldsymbol{x}_n(t_{n+l}\text{+}0) &= \Psi_{n+l}\boldsymbol{x}_n(t_{n+l-1}\text{+}0) + \boldsymbol{g}u_{l,n}, \quad l = 1,2,3,4, \end{aligned} \quad (27)$$

where we have set

$$\Psi_{n+l} = \Phi(t_{n+l}\text{+}0, t_{n+l}\text{-}0)\Phi(t_{n+l}\text{-}0, t_{n+l-1}\text{+}0), \quad l = 1,2,3,4, \quad (28)$$

and used our choice $u_{0,n} = 1$ and the initial state $\boldsymbol{x}(t_n\text{-}0) = \boldsymbol{0}$. By these recurrence formulae, we can write the terminal state as follows:

$$\begin{aligned} \boldsymbol{x}_n(t_{n+4}\text{+}0) = {} & \Psi_{n+4}\Psi_{n+3}\Psi_{n+2}\Psi_{n+1}\boldsymbol{g} \\ & + \Psi_{n+4}\Psi_{n+3}\Psi_{n+2}\boldsymbol{g}u_{1,n} \\ & + \Psi_{n+4}\Psi_{n+3}\boldsymbol{g}u_{2,n} \\ & + \Psi_{n+4}\boldsymbol{g}u_{3,n} \\ & + \boldsymbol{g}u_{4,n}. \end{aligned} \quad (29)$$

Then we can determine u_1, u_2, u_3, and u_4 that makes the terminal state $\boldsymbol{x}_n(t_{n+4}\text{+}0)$ be zero by

$$\begin{bmatrix} u_{1,n} \\ u_{2,n} \\ u_{3,n} \\ u_{4,n} \end{bmatrix} = -\left[\Psi_{n+4}\Psi_{n+3}\Psi_{n+2}\boldsymbol{g} \;\; \Psi_{n+4}\Psi_{n+3}\boldsymbol{g} \;\; \Psi_{n+4}\boldsymbol{g} \;\; \boldsymbol{g}\right]^{-1} \Psi_{n+4}\Psi_{n+3}\Psi_{n+2}\Psi_{n+1}\boldsymbol{g}. \quad (30)$$

Existence of the inverse matrix is equivalent to the complete controllability of the sampled-data system with the impulse control u_n input. We do not have the condition in a simpler form due to the complication caused by time-varying dynamics and non-uniform sampling. Even the uniform sampling case is yet to be investigated.

For the numerical evaluation of y_n, we first compute the states $\{\boldsymbol{x}_n(t_{n+l}\text{+}0)\}_{l=0}^3$ by (27) from $\{u_{l,n}\}_{l=0}^4$. Then we can evaluate y_n by

$$y_n(t) = \begin{cases} 0, & t \le t_n \\ \boldsymbol{h}\Phi(t, t_{n+l}\text{+}0)\boldsymbol{x}_n(t_{n+l}\text{+}0), & \\ \qquad t_{n+l} < t \le t_{n+l+1}, (l = 0,1,2,3) & \\ 0, & t_{n+4} \le t \end{cases} \quad (31)$$

and

$$\boldsymbol{h}\Phi(t, t_{n+l}\text{+}0) = \begin{cases} \begin{bmatrix} 1 & t - t_{n+l} & \frac{(t-t_{n+l})^2}{2} & \frac{(t-t_{n+l})^3}{6} \end{bmatrix} & \text{if } p_{n+l} = 0 \\ \begin{bmatrix} 1 & t - t_{n+l} & \frac{\cosh(p_{n+l}(t-t_{n+l}))-2}{p_{n+l}^2} & \\ & & \frac{\sinh(p_{n+l}(t-t_{n+l}))-2p_{n+l}(t-t_{n+l})}{p_{n+l}^3} \end{bmatrix} & \text{if } p_{n+l} > 0 \end{cases} \quad (32)$$

which follow from (12), (16), and the continuity of y_n over the entire domain.

4. Numerical examples

Test data were prepared by concatenating a sampled smooth curve and a sampled polygonal line. Their interpolation was computed as a linear combination of the locally supported splines in tension.

The cubic spline interpolation (equivalent to the case $p(t) \equiv 0$) is shown in Fig. 1. The cursive part is reproduced in a good shape but the polygonal part suffers from inter-sample vibration. The linear spline interpolation (equivalent to the case $p(t) \to \infty$) in Fig. 2 behaves in the opposite way. Reproduction of the polygonal part is perfect but there is no smoothness. Interpolation by a spline in constant tension ($p(t) \equiv 10$) in Fig. 3 provides a good compromise between the cubic and linear spline interpolation. It is fairly smooth and has less vibration.

Some may say that the cursive part is not smooth enough and rather polygonal in Fig. 3. In this case, we can obtain a better interpolation by varying the tension in time. Figure 4 is an interpolation by a spline in piecewise constant tension. Higher tensions are imposed on the polygonal part to suppress the vibration. The interpolation is kept smooth elsewhere. The locally supported splines used to construct this curve are plotted in Fig. 5 where the plots are vertically scaled to have a common peak value.

5. Conclusions

Locally supported splines in tension were constructed where the tension is constant within each sampling interval and variable at the sampling points. They will hopefully contribute to the variety of curve drawing modules in the graphical design tools. Another application may be image enlargement tools which allow users to put higher tension manually at the portions where they want to suppress ringing effects.

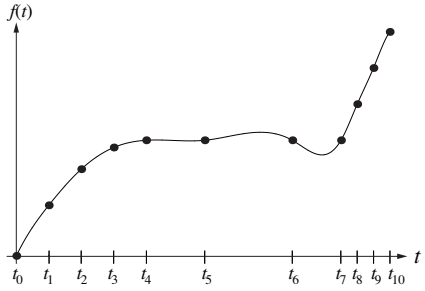


Figure 1: Interpolation by a cubic spline ($p(t) \equiv 0$).

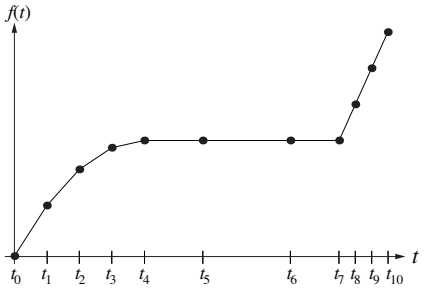


Figure 2: Interpolation by a linear spline ($p(t) \to \infty$).

References:

[1] J. H. Ahlberg, E. N. Nilson and J. L. Walsh. *The Theory of Splines and Their Applications*. Academic Press, London, 1967.

[2] M. N. Benbourhim and A. Bouhamidi. Approximation of vector fields by thin plate splines with tension. *J. Approx. Theory*, 136:198–229, 2005.

[3] C. de Boor. Best approximation properties of spline functions of odd degree. *J. Math. Mech.*, 12:747–750, 1963.

[4] Y. C. Ho, R. E. Kalman and K. S. Narendra. Controllability of linear dynamical systems. *Contrib. Diff. Eqs.*, 1:189–213, 1963.

[5] J. C. Holladay. Smoothest curve approximation. *Math. Tables and Aids to Comput.*, 11:223–243, 1957.

[6] R. E. Kalman. A new approach to linear filtering and prediction problems. *Trans. ASME*, 82(Series D):35–45, 1960.

[7] H. Mitasova and L. Mitas. Interpolation by regularized spline with tension: I. theory and implementation. *Mathematical Geology*, 25:641–655, 1993.

[8] A. Polyakov and V. Brusentsev. *Graphics Programming with GDI+ & DirectX*. A-List Publishing, Wayne, PA, 2005.

[9] I. J. Schoenberg. On interpolation by spline functions and its minimal properties. In P. L. Butzer and J. Korevaar, editors, *On Approximation Theory*, pages 109–129, June 1964.

[10] D. G. Schweikert. An interpolation curve using a spline in tension. *J. Math. Phys.*, 45:312–317, 1966.

[11] E. D. Sontag. *Mathematical Control Theory*. Springer, New York, 1990.

[12] M. Unser and T. Blu. Cardinal exponential splines: Part I—Theory and filtering algorithms. *IEEE Transactions on Signal Processing*, 53(4):1425–1438, April 2005.

[13] M. Unser. Cardinal exponential splines: Part II—Think analog, act digital. *IEEE Transactions on Signal Processing*, 53(4):1439–1449, April 2005.

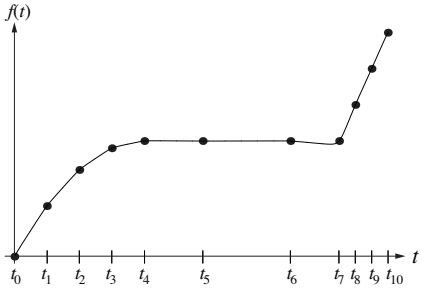


Figure 3: Interpolation by a spline in constant tension ($p(t) \equiv 10$).

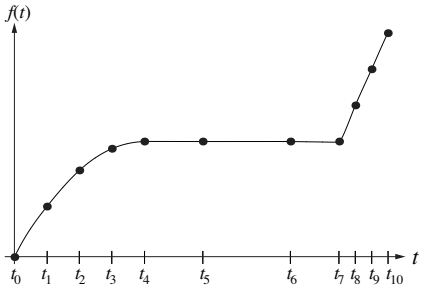


Figure 4: Interpolation by a spline in piecewise constant tension ($p(t) = 0$ for the cursive part ($t < t_4$), $p(t) = 10$ for the straight parts ($t_4 \leq t < t_6$ and $t_7 \leq t$), and $p(t) = 30$ for the breaking part ($t_6 \leq t < t_7$)).

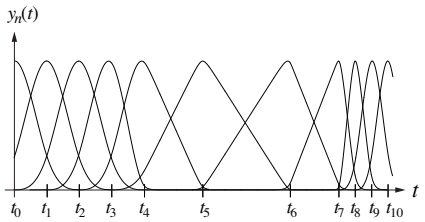


Figure 5: Locally supported splines used to construct the curve in Fig. 4.

The Effect of Sampling Frequency on a FFT Based Spectral Estimator

Saeed Ayat

Payame Noor University, Najafabad, Iran.
dr.ayat@pnu.ac.ir

Abstract:

This paper reviews the effect of sampling frequency on a FFT-based spectral estimator. In signal processing applications usually a fix window size is used for obtaining the current frame spectral.
For an application like speech enhancement this accuracy of this estimation has a great influence in the quality of the system, because listener feeling is very important in this subject.
In our proposed method we divided the well-known spectral subtraction method in two phases. Then by using different frame sizes that we used in these two phases the overall quality of the system has increased in different sampling frequencies.

1. Introduction

One of the first methods introduced for speech enhancement is spectral subtraction. Till now, different versions of spectral subtraction have been proposed to increase the performance of this method, for example [1, 2, 3].
Despite of its high noise removal, it can cause an annoying noise called musical noise and hence it can reduce overall quality. Musical noise is produced because, we don't have the needed spectra exactly, so we have to use their estimations.
In signal processing applications usually a fix window size is used for obtaining the current frame spectral. As we know if the frame length is L the frequency resolution in Fourier spectral analysis is Fs/L. For example if F_s=11025Hz and L=256 then Fs/L is 43Hz and this resolution may not be enough for speech signal.
As we know a clean speech signal consists of some sections that have speech and some others that have no speech and we call them silences.
In a noisy speech signal these silence sections have only noise and other sections have noisy speech signals. If the noise is stationary we can estimate its spectrum in the noise sections.
In spectral subtraction method, after framing the noisy speech signal we use a silence detector or a voice activity detector for separating noisy speech frames and noise frames.
After that with applying, FFT we have the spectrum of each frame. By calculating the average of the noise frames spectra we have estimation for noise spectra. Now with subtracting this estimation of noise spectrum from the spectrum of each noisy speech frames we can achieve enhanced speech signal.
The paper is organized as follows: In section 2 we have a review on spectral subtraction method. In section 3 we proposed our method and in section 4 we present the simulation results.

2. Spectral Subtraction

There are many different versions for spectral subtraction. In a generalized spectral subtraction [4] we have:

$$\left|\hat{S}(w)\right| = \max\left\{\left(\left|S(w)\right|^{\alpha} - \beta\left|N(w)\right|^{\alpha}\right)^{\frac{1}{\alpha}}, \gamma\left|N(w)\right|\right\} \quad (1)$$

Where $\left|S(w)\right|, \left|N(w)\right|$ and $\left|\hat{S}(w)\right|$ are magnitude spectrum of noisy speech, estimation of noise and enhanced speech. β is the oversubtraction factor and γ is spectral floor. Both β and γ are adjusted to improve the quality of enhanced speech.
By the assuming that the noise is stationary, a good estimation can be resulted by computing the average of the noise in silence frames spectra. We called such average $\left|\overline{W}(w)\right|$.
In presence of nonstationary noises, an adaptation technique can be used. Given an initial value $\left|\overline{W}_0(w)\right|$, if the current frame is silence, $\left|\overline{W}_m(w)\right|$ is updated using this equation:

$$\left|\overline{W}_m(w)\right| = (1-f)\left|\overline{W}_{m-1}(w)\right| + f\left|Y_m(w)\right| \quad (2)$$

In this formula $\left|Y_m(w)\right|$ is the spectrum of current silence frame and f is a coefficient called forgetting factor. This factor is changed depending on the noise changing rate.
The main problem of spectral subtraction method is the production of musical noise. Musical noise is produced because we don't have the exact spectrum of the noise signal.

3. Proposed Method

In our method that estimates the spectrum better than the basic averaging method, after separating speech and silence frames in the noisy signal with a basic analysis frame, we can increase the analysis frame length until it covers all the current silence frames. As in periodogram estimator technique the accuracy improves by increasing the number of signal samples, by using this adaptive analysis frame length we can have a better spectral estimation for noise and noisy signal and so the system can produce a better enhanced signal with less musical noise.

As we know if the frame length is L the frequency resolution in Fourier spectral analysis is Fs/L. For example if F_s=11025Hz and L=256 then Fs/L is 43Hz and this resolution may not be enough for speech signal.

In our method we first apply a SAD algorithm with L=256 and L/2=128 points overlap to detect the silence frames. Now we can increase the analysis frame length until it covers all the current silence frames. By this method we have larger window length and hence better frequency resolution. If we have several silence areas with the new frame length, the average of them is the overall noise spectrum.

By applying such method we have better noise spectrum estimation with less musical noises.

In section 4 we give experimental results that confirm this improvement clearly.

4. Simulation Results

In this section we explain our simulation. The speech signal that used for these tests was chosen from TIMIT data base and was pronounced with a female speaker. Then this sentence converted to different sampling frequencies by cool edit software. All these sentences degraded by additive Gaussian white noise, so we can have the noisy signal in required SNR, here 5dB.

For evaluating our method we calculate SNR improvement as below.

If $s(n)$ is the clean speech, y(n) the noisy, $\hat{s}(n)$ the enhanced signal and w(n) the noise then we have:

$$y(n) = s(n) + w(n) \tag{3}$$

and the SNR improvement is computed as follows[5]:

$$SNR_{imp} = SNR_{out} - SNR_{in} \tag{4}$$

In which SNR_{in} and SNR_{out} are the SNRs for noisy and enhanced:

$$SNR_{in} = 10\log_{10}\frac{\sum s^2(n)}{\sum (y(n)-s(n))^2} \tag{5}$$

$$SNR_{out} = 10\log_{10}\frac{\sum s^2(n)}{\sum (\hat{s}(n)-s(n))^2} \tag{6}$$

In this experiment a listener listens to the enhanced signal and increases β until the musical noise appeares in the enhanced signal. At this point, β and SNR improvement is recorded. This is done for SNR equal to 5dB and different frame lengths with 256, 512, 1024 and 2048 samples.

This test was evaluated for different sampling frequencies equal to 8000Hz, 11025 Hz and 16000Hz. α is fixed to 1.0 and γ to 0.0. Note that the frame length is 256 in silence detection step.

Tables 1 to 3 show the results for β and SNR improvement at the appearance of musical noise in the enhanced signal for tested SNRs.

L	256	512	1024	2048
SNR_{imp}	0.8	1.0	1.4	1.44
β	0.1	0.15	0.25	0.45

Table 1: β and SNR improvement at the start of musical noise (Fs=8000Hz)

L	256	512	1024	2048
SNR_{imp}	1.0	1.8	2.3	3.1
β	0.15	0.3	0.5	0.9

Table 2: β and SNR improvement at the start of musical noise (Fs=11025Hz)

L	256	512	1024	2048
SNR_{imp}	1.2	2.3	3.1	3.8
β	0.2	0.5	0.7	1.0

Table 3: β and SNR improvement at the start of musical noise (Fs=16000Hz)

As we can see the SNR improvement is better for longer frame lengths in all different sampling frequency rates. This show that the musical noise arises from inaccurate noise estimation and reduces as the frame length increases, and this result is true for different sampling frequencies.

So with a greater frame length, we can choose a greater β without production of musical noise and by increasing it we can have less noise in the enhanced signal and then achieve more SNR improvement, too.

5. Conclusions

In this paper we studied the effect of sampling frequency on a FFT-based spectral estimator. We also proposed an improved spectral subtraction method by increasing the accuracy of spectral estimator.
This adaptive estimator can give better spectral estimation by increasing the analysis frame length that achieves in silence regions.
In this method for separating silence frames we use a basic analyzing frame and for estimation the spectrum we use an adaptive frame length that can increase until it covers all current silence region. By this method we could have a better spectral estimation for noise and noisy signal and so the system can produce a better enhanced signal with less musical noise.

References:

[1] S. F. Boll, "Suppression of Acoustic Noise in Speech using Spectral Subtraction," IEEE Trans. Acoustics, Speech and Signal processing, vol. ASSP-27, No. 2, pp. 113-120, 1977.

[2] H. Hu, F. Kuo, H. Wang, "Supplementary Schemes to Spectral Subtraction for Speech Enhancement," Speech Communication, 2002.

[3] H. Gustafsson, S. Nordholm, "Speech Subtraction using Reduced Delay Convolution and Adaptive Averaging", *IEEE Trans. Speech and Audio Processing*, vol. 9, No. 8, pp. 799-807, 2001.

[4] J. S. Lim, A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech", Proceedings of the IEEE, vol. 67, 1972.

[5] S. Ayat, "Enhanced Human-Computer Speech Interface Using Wavelet Computing", IEEE International Conference on Virtual Environments, Human-Computer Interfaces and Measurement Systems, 2008.

Nonlinear Locally Adaptive Wavelet Filter Banks

Gerlind Plonka (1) and Stefanie Tenorth (1)

(1) Department of Mathematics, University of Duisburg-Essen, 47048 Duisburg, Germany.
gerlind.plonka@uni-due.de, stefanie.tenorth@uni-due.de

Abstract:

In this paper we introduce a new construction of nonlinear locally adaptive wavelet filter banks by connecting the lifting scheme with the idea of image smoothing by nonlinear diffusion methods.

1. Introduction

A crucial problem in data analysis is to construct efficient low-level representations, thereby providing a precise characterization of features which compose it, such as edges and texture components. Fortunately, in many relevant applications, the components of given multidimensional data are not independent, and the strong correlation between neighboring data points can be suitably exploited. In the two-dimensional case, tensor-product wavelets are not optimal for representing geometric structures because their support is not adapted to directional geometric properties.

Instead of choosing a priori a basis or a frame to approximate the image, one can try to adopt the approximation scheme to the image geometry. Within the last years, different approaches have been developed in this direction, see e.g. [1, 4, 5, 7, 10, 12, 13]. In particular, the construction of non-linear filter banks by the lifting scheme has been proposed already in [4, 8]. Since that time, there have been different attempts to construct adaptive and directional lifting based, invertible transforms for sparse image representation, see [2, 5, 6, 9, 12]. The lifting scheme for representation of wavelet filter banks has originally been suggested and analyzed by Sweldens [16]. It provides a flexible tool for the construction of new nonlinear wavelet filter banks. The main feature of lifting is that it provides an entirely spatial-domain interpretation of the transform.

Besides wavelet shrinkage, other approaches like regularization techniques and PDE-based methods (as nonlinear diffusion) have been shown to be powerful tools in signal and image restoration in image processing, e.g., for denoising purposes. In particular, the choice of nonlinear diffusion filters leads to impressive results by removing insignificant, small-scale variations while preserving important features such as discontinuities [3, 11, 17, 18]. In [15], certain connections between explicit discrete one-dimensional schemes for non-linear diffusion and shift-invariant Haar wavelet shrinkage have been established.

In this paper we wish to construct a new invertible nonlinear wavelet filter bank by connecting the two concepts of the lifting scheme and the discrete nonlinear diffusion. The main goal is to adapt the local geometry of images suitably, in order to obtain highly efficient sparse image representations.

2. Lifting and Nonlinear Diffusion

2.1 The Lifting Scheme

The typical lifting scheme consists of three steps: Split, Predict and Update.

1. Split. Usually, in this step, the given data is split into even and odd components. Let $N \in \mathbb{N}$ be of the form $N = 2^l\, r$ with $l,\, r \in \mathbb{N}$. For a given digital image of the form $a = (a(i,j))_{i,j=0}^{N-1} \in \mathbb{R}^{N\times N}$, we split the data into the following two sets of equal size,

$$\begin{aligned} a^e &:= (a_{i,j})_{i,j=0,i+j \text{ even}}^{N-1}, \\ a^o &:= (a_{i,j})_{i,j=0,i+j \text{ odd}}^{N-1}, \end{aligned}$$

and we denote the components of a^e and a^o by $a^e_{i,j}$ and $a^o_{i,j}$, respectively. The data sets a^e and a^o split the image a like a checkerboard.

2. Predict. The goal of the prediction step is to find a good approximation $\tilde{a}^o$ of the data a^o of the form

$$\tilde{a}^o = P_1(a^o) + P_2(a^e).$$

Here P_1 and P_2 can be nonlinear operators. Afterwards, we consider the residual

$$d^o := a^o - \tilde{a}^o = a^o - (P_1(a^o) + P_2(a^e)).$$

We have to assume that the mapping $(a^e, a^o) \mapsto (a^e, d^o)$ is invertible, i.e., the operator $I - P_1$ needs to be invertible for arbitrary data a^o. The operators P_1 and P_2 need to be chosen such that the residual d^o is very small.

3. Update. In the third step, we aim to find a smoothed approximation of the data a^e that can be regarded as a low-pass filtered and subsampled version of the original image a. The general update has the form

$$\tilde{a}^e := U_1(d^o) + U_2(a^e)$$

with (possibly nonlinear) operators U_1 and U_2, where we again want to assume the invertibility of the mapping

$(a^e, d^o) \mapsto (\tilde{a}^e, d^o)$, i.e., U_2 is assumed to be invertible such that

$$a^e = U_2^{-1}(\tilde{a}^e - U_1(d^o)).$$

The complete scheme is illustrated in Figure 1.

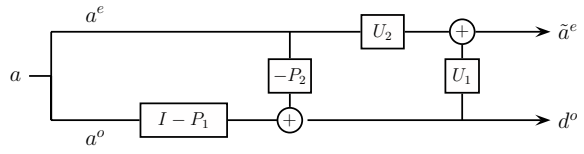


Figure 1: Illustration of the nonlinear filter bank using the lifting scheme.

2.2 Nonlinear Diffusion

The nonlinear diffusion has been shown to be a very successful model for image denoising. For $\Omega = (0, N_1) \times (0, N_1)$ we consider the diffusion equation

$$\frac{\partial u}{\partial t} = \text{div}\left(g(|\nabla u|)\,\nabla u\right) \quad \text{on } \Omega \times (0, \infty) \qquad (1)$$

with a given noisy image a as initial state

$$u(x, 0) = a(x), \quad x \in \overline{\Omega}$$

and with Neumann boundary conditions $\frac{\partial u}{\partial n} = 0$ on $\partial\Omega$. Here, $\nabla u = (u_{x_1}, u_{x_2})^T = (\partial u/\partial x_1, \partial u/\partial x_2)^T$ denotes the gradient of u, and $|\nabla u| := \sqrt{u_{x_1}^2 + u_{x_2}^2}$.
The time t in (1) is a scale parameter. Increasing t corresponds to stronger filtering. The diffusivity function $g(|\nabla u|)$ is a non-negative function that determines the amount of diffusion. It is decreasing in $|\nabla u|$ in order to ensure that strong edges are hardly blurred by the diffusion filter while small variations (noise) are smoothed much stronger. Frequently used bounded diffusivities are the Perona-Malik diffusivity

$$g(x) := \frac{1}{1 + x^2/\lambda^2},$$

or the Weickert diffusivity

$$g(x) := \begin{cases} 1 & x = 0, \\ 1 - \exp\left(\frac{-3.315}{(x/\lambda)^4}\right) & x > 0, \end{cases}$$

see [14, 17]. One may also take a "robust" diffusivity of the form

$$g(x) := \begin{cases} 1 & 0 \le x < \theta, \\ 0 & |x| \ge \theta, \end{cases}$$

as it has been used in [14] with a suitably chosen threshold θ.
Replacing $g(|\nabla u|)$ by $g(|\nabla u_\sigma|)$, where u_σ denotes the slightly smoothed image by convolution with the Gaussian kernel, $u_\sigma := K_\sigma \star u$, existence and uniqueness of a solution of (1) have been shown in [3].
For application of the diffusion approach to digital images we follow [11] and replace (1) by the following slightly modified equation

$$\frac{\partial u}{\partial t} = \partial_{x_1}(g(|\partial_{x_1} u|)\,\partial_{x_1} u) + \partial_{x_2}(g(|\partial_{x_2} u|)\,\partial_{x_2} u).$$

We use a discretization of the form

$$\begin{aligned} \frac{u_{ij}^{k+1} - u_{ij}^k}{\tau} &= g(|u_{i+1,j}^k - u_{i,j}^k|)(u_{i+1,j}^k - u_{i,j}^k) \\ &\quad - g(|u_{i,j}^k - u_{i-1,j}^k|)(u_{i,j}^k - u_{i-1,j}^k) \\ &\quad + g(|u_{i,j+1}^k - u_{i,j}^k|)(u_{i,j+1}^k - u_{i,j}^k) \\ &\quad - g(|u_{i,j}^k - u_{i,j-1}^k|)(u_{i,j}^k - u_{i,j-1}^k), \end{aligned} \qquad (2)$$

where $u_{i,j}^0 := a_{ij}$ for $i, j = 0, \ldots, N-1$. Here, k denotes the iteration step and τ is the step size of time discretization. In our numerical examples we will use the step size $\tau = 1/4$.

3. The Nonlinear Diffusion Filter Bank

Now we want to apply the nonlinear diffusion filter for the construction of prediction and update operators in the lifting scheme, in order to obtain a new sparse representation of images. The nonlinear filter bank should satisfy the following demands.
1. For linear (bivariate) polynomials, the residual d^o found in the prediction step should vanish. This condition is equivalent with two vanishing moments of the high-pass filter in a wavelet filter bank.
2. Near discontinuities (edges) of u, the residual d^o should remain small.
3. The data $\tilde{a}^e$ should be a suitable (downsampled) approximation of the image a with good low-pass filter properties in smooth areas of a and without blurring of edges.

3.1 Choice of the Prediction Operator

Using equation (2) with the notations $a_{i,j}^o := u_{i,j}^0$, $\tilde{a}_{i,j}^o := u_{i,j}^1$ for $i+j$ odd, and $a_{i,j}^e = u_{i,j}^0$ for $i+j$ even, we obtain

$$\begin{aligned} \tilde{a}_{i,j}^o \;=\; & a_{i,j}^o + \tau\left[g(|a_{i+1,j}^e - a_{i,j}^o|)(a_{i+1,j}^e - a_{i,j}^o)\right. \\ & + g(|a_{i-1,j}^e - a_{i,j}^o|)(a_{i-1,j}^e - a_{i,j}^o) \\ & + g(|a_{i,j+1}^e - a_{i,j}^o|)(a_{i,j+1}^e - a_{i,j}^o) \\ & \left. + g(|a_{i,j-1}^e - a_{i,j}^o|)(a_{i,j-1}^e - a_{i,j}^o)\right]. \end{aligned}$$

A prediction could now be of the form

$$\begin{aligned} & d_{i,j}^o = a_{i,j}^o - \tilde{a}_{i,j}^o \\ & = -\tau\Big[\sum_{\substack{\mu,\nu=-1 \\ |\mu|+|\nu|=1}}^{1} g(|a_{i+\mu,j+\nu}^e - a_{i,j}^o|)(a_{i+\mu,j+\nu}^e - a_{i,j}^o)\Big]. \end{aligned}$$

Unfortunately, with this coice of prediction the desired invertibility of the mapping $(a^e, a^o) \mapsto (a^e, d^o)$ is not guaranteed since the nonlinear diffusion g depends on the data $a_{i,j}^o$. Therefore, we replace the values $a_{i,j}^o$ that are used for the computation of the function values of g by the median of its four direct neighbors,

$$a_{i,j}^o \approx \text{median}\,\{a_{i,j+1}^e, a_{i,j-1}^e, a_{i+1,j}^e, a_{i-1,j}^e\} := \text{med}\, a_{i,j}^o.$$

A normalization with

$$g_{ij} := \sum_{\substack{\mu,\nu=-1 \\ |\mu|+|\nu|=1}}^{1} g(|a_{i+\mu,j+\nu}^e - \text{med}\, a_{i,j}^o|)$$

now yields the prediction

$$d^o_{i,j} := \frac{-\tau}{g_{ij}} \sum_{\substack{\mu,\nu=-1 \\ |\mu|+|\nu|=1}}^{1} g(|a^e_{i+\mu,j+\nu} - \operatorname{med} a^o_{i,j}|)(a^e_{i+\mu,j+\nu} - a^o_{i,j})$$

$$= \tau a^o_{i,j} - \frac{\tau}{g_{ij}} \sum_{\substack{\mu,\nu=-1 \\ |\mu|+|\nu|=1}}^{1} g(|a^e_{i+\mu,j+\nu} - \operatorname{med} a^o_{i,j}|)\, a^e_{i+\mu,j+\nu}.$$

Now,the invertibility of the prediction is ensured for $\tau > 0$ and we have

$$a^o_{i,j} = \frac{d^o_{i,j}}{\tau} + \frac{1}{g_{ij}} \sum_{\substack{\mu,\nu=-1 \\ |\mu|+|\nu|=1}}^{1} g(|a^e_{i+\mu,j+\nu} - \operatorname{med} a^o_{i,j}|) a^e_{i+\mu,j+\nu}.$$

Observe that the term g_{ij} is positive for all i, j if we take Perona-Malik diffusivity or Weickert diffusivity. At the boundary of the image, where not all four neighbors of a data point are given, we slightly change the operator and use only the three available neighbors in the sum (or even only two neighbors at a vertex). Because of the normalization with the (correspondingly defined constants g_{ij}) the properties of the prediction operator will not change.

3.2 Choice of the Update Operator

As update operator we simply apply a linear operator of the form

$$\tilde{a}^e_{i,j} = \sqrt{2} a^e_{i,j} + \tfrac{1}{4}(d^o_{i+1,j} + d^o_{i-1,j} + d^o_{i,j+1} + d^o_{i,j-1}).$$

Invertibility is obviously satisfied and we find

$$a^e_{i,j} = \tfrac{1}{\sqrt{2}} \left(\tilde{a}^e_{i,j} - \tfrac{1}{4}(d^o_{i+1,j} + d^o_{i-1,j} + d^o_{i,j+1} + d^o_{i,j-1}) \right).$$

At the boundary, where $a^e_{i,j}$ has only three neighbors, we slightly change the operator. For example, for $0 < i < N-1$ and $j = 0$, we take

$$\tilde{a}^e_{i,0} := \sqrt{2} a^e_{i,0} + \tfrac{1}{3}(d^o_{i+1,0} + d^o_{i-1,0} + d^o_{i,1}),$$

etc.. Analogously, at vertices, only two neighbors are taken into account.
Observe that the low-pass filtered values $\tilde{a}^e_{i,j}$ are amplified by $\sqrt{2}$ here (as it is usual also for orthogonal wavelet filter banks).

3.3 Iterative Application of the Filter Bank

In order to obtain a suitable sparse representation of the digital image a, we now iteratively apply the nonlinear filter bank described above, and we use a hard threshold procedure to suppress small residual values $d^o_{i,j}$.
After the first application of the filter bank, the (small) residual data $d^o_{i,j}$, $i, j = 0, \ldots, N-1$, $i + j$ odd, are stored and we consider only the $N^2/2$ values $\tilde{a}^e_{i,j}$, $i, j = 0, \ldots, N-1$, $i + j$ even. For a second application of the filter bank to $\tilde{a}^e_{i,j}$, we rename these data by $a^{(1)}_{k,l} := \tilde{a}^e_{k-l,k+l}$, where $k = 0, \ldots, N-1$ and $l = -\min\{k, N-1-k\}, \ldots, \min\{k, N-1-k\}$, and apply the filters now to this data set, etc..
As usual, the complete procedure involves the following three steps. First, we decompose the image by an iterative application of the diffusion filter bank. Secondly, we apply the shrinkage function

$$S_\theta(x) := \begin{cases} x & |x| \geq \theta, \\ 0 & |x| < \theta, \end{cases}$$

to the residual coefficients. In our numerical experiments we will take a level-independent threshold θ. Finally, we reconstruct the image with the modified residual coefficients.

4. Properties of the Diffusion Filter Bank

We can show the following

Theorem 1.
Let g be a diffusivity function satisfying $0 < g(|x|) \leq 1$ for $x \in \mathbb{R}$. The diffusion filter bank determined in Section 3 reproduces linear polynomials.

Proof. We consider a linear polynomial of the form

$$a(x_1, x_2) = a_0 + b_0 x_1 + c_0 x_2, \quad a_0, b_0, c_0 \in \mathbb{R}.$$

Let the digital image now be given by

$$a_{i,j} = a(ih, jh) = a_0 + b_0\, ih + c_0\, jh.$$

Then we obtain for data that are not at the boundary

$$\begin{aligned} \operatorname{med} a^o_{i,j} &= \text{median}\,\{a_0 + b_0(i-1)h + c_0 jh, a_0 + \\ &\quad b_0(i+1)h + c_0 jh, a_0 + b_0 ih + c_0(j-1)h, \\ &\quad a_0 + b_0 ih + c_0(j+1)h\} \\ &= a_0 + b_0 ih + c_0 jh + \\ &\quad \text{median}\,\{-b_0 h, b_0 h, -c_0 h, c_0 h\} \\ &= a_0 + b_0 ih + c_0 jh = a^o_{i,j} \end{aligned}$$

and

$$\begin{aligned} d^o_{i,j} &= \frac{-\tau}{g_{ij}} \sum_{\substack{\mu,\nu=-1 \\ |\mu|+|\nu|=1}}^{1} g(|a^e_{i+\mu,j+\nu} - a^o_{i,j}|)(a^e_{i+\mu,j+\nu} - a^o_{i,j}) \\ &= \frac{-\tau}{g_{ij}} \big[g(b_0 h)(a^e_{i+1,j} + a^e_{i-1,j} - 2a^o_{i,j}) \\ &\qquad + g(c_0 h)(a^e_{i,j+1} + a^e_{i,j-1} - 2a^o_{i,j}) \big] \\ &= 0. \end{aligned}$$

Hence the prediction operator leads to $d^o_{i,j} = 0$ and the update yields $\tilde{a}^e_{i,j} = \sqrt{2} a^e_{i,j}$ for all i, j with $i + j$ even. □

Further, one can show in case studies, that the proposed filter bank behaves well at vertical, horizontal and diagonal edges, i.e., the obtained residual values using the nonlinear prediction operator remain to be small.

5. Numerical Results

We apply the above described nonlinear diffusion filter bank in order to achieve sparse image representations. In the experiment, we consider the monarch image. We use the Perona-Malik diffusivity with $\lambda = 28$ and with $\tau = 0.25$. We apply 8 levels of the nonlinear filter bank,

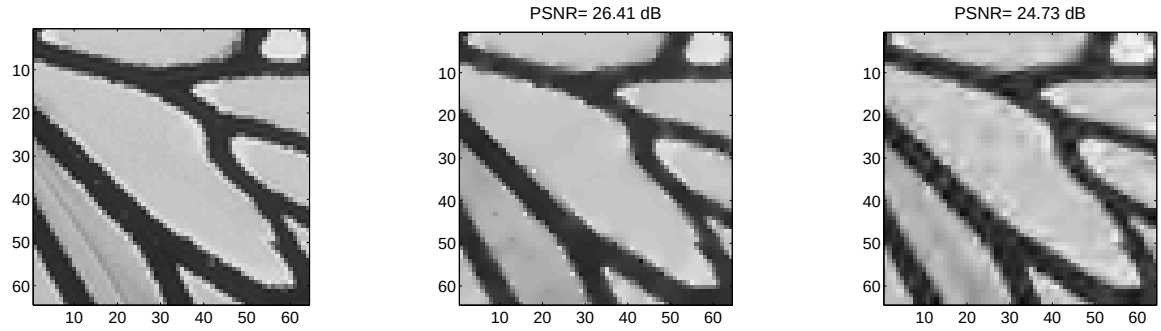


Figure 2: Original image Monarch (left), sparse image representation with 449 coefficients using the proposed nonlinear diffusion filter bank (middle) and the biorthogonal filter bank with 7-9 filter (right).

i.e., there will remain 16 low-pass coefficients. For thresholding we use the hard shrinkage function with $\theta = 13$. In Figure 2(left), we present the original image. Figure 2(middle) shows the obtained compressed image with 449 remaining coefficients using the new diffusion filter bank. For comparison, we apply 8 decomposition levels of the two-dimensional biorthogonal wavelet shrinkage with the $7-9$ filter with the same number of 449 remaining nonzero coefficients, see Figure 2(right). As we can see, the nonlinear filter bank not only gives an optically better result but also achieves a better PSNR value (26.41 dB) while the biorthogonal filter bank achieves a PSNR of 24.73 dB. We remark that our method is especially designed for constructing efficient low-level representations and does not work well for image denoising.

6. Acknowledgement

The research in this paper is supported by the project PL 170/13-1 of the German Research Foundation (DFG). This is gratefully acknowledged.

References:

[1] F. Arandiga, A. Cohen, R. Donat, N. Dyn. Interpolation and approximation of piecewise smooth functions. *SIAM J. Numer. Anal.* 43:41–57, 2005.

[2] N.V. Boulgouris, D. Tzovaras, and M.G. Strintzis. Lossless image compression based on optimal prediction, adaptive lifting, and conditional arithmetic coding. *IEEE Trans. Image Process.* 10:1–14, 2001.

[3] F. Catté, P.-L. Lions, J.-M. Morel, and T. Coll, Image selective smoothing and edge detection by nonlinear diffusion. *SIAM J. Numer. Anal.* 29:182–193, 1992.

[4] R.L. Claypoole, G.M. Davis, W. Sweldens, and R.G. Baraniuk. Nonlinear wavelet transforms for image coding via lifting. *IEEE Trans. Image Process.* 12:1449–1459, 2003.

[5] A. Cohen and B. Matei. Compact representation of images by edge adapted multiscale transforms. In *Proc. IEEE Int. Conf. on Image Process. (ICIP)*, Thessaloniki, pages 8–11, 2001.

[6] W. Ding, F. Wu, X, Wu, S. Li, and H. Li. Adaptive directional lifting-based wavelet transform for image coding. *IEEE Trans. Image Process.* 16:416–427, 2007.

[7] D.L. Donoho. Wedgelets: Nearly minimax estimation of edges. *Ann. Stat.* 27:859–897, 1999.

[8] F.J. Hampson and J.-C. Pesquet. A nonlinear subband decomposition with perfect reconstruction. *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, pages 1523–1526, 1996.

[9] H.J.A.M. Heijmans, B. Pesquet-Popescu, G. Piella. Building nonredundant adaptive wavelets by update lifting. *Appl. Comput. Harmon. Anal.* 18:252–281, 2005.

[10] S. Mallat. Geometrical grouplets. *Appl. Comput. Harmon. Anal.* 26:161–180, 2009.

[11] P. Perona and J. Malik. Scale space and edge detection using anisotropic diffusion. Proc. IEEE Computer Society Workshop on Computer Vision, IEEE Computer Society Press, pages 16–22, 1987.

[12] G. Piella, B. Pesquet-Popescu, H.J.A.M. Heijmans, and G. Pau. Combining seminorms in adaptive lifting schemes and applications in image analysis and compression. *J. Math. Imaging Vis.* 25:203–226, 2006.

[13] G. Plonka. The easy path wavelet transform: A new adaptive wavelet transform for sparse representation of two-dimensional data. *Multiscale Model. Simul.* 7:1474–1496, 2009 .

[14] G. Plonka and J. Ma. Convergence of an iterative nonlinear scheme for denoising of piecewise constant images. *Int. J. Wavelets Multiresolut. and Inf. Process.* 5:975–995, 2007.

[15] G. Steidl, J. Weickert, T. Brox, P. Mrázek, M. Welk. On the equivalence of soft wavelet shrinkage, total variation diffusion, total variation regularization, and sides. *SIAM J. Numer. Anal.* 42/2:686–713, 2004.

[16] W. Sweldens. The lifting scheme: A construction of second generation wavelets. *SIAM J. Math. Anal.* 29:511–546, 1997.

[17] J. Weickert. *Anisotropic Diffusion in Image Processing*. Teubner, Stuttgart, 1998.

[18] M. Welk, G. Steidl, and J. Weickert. A four–pixel scheme for singular differential equations. In R. Kimmel, N. Sochen, J. Weickert, editors, *Scale-Space and PDE Methods in Computer Vision.* Lecture Notes in Computer Science, Springer, Berlin, pages 610–621 (2005).